

基于 BERT 的民间文学文本预训练模型

陶慧丹^{1,2}, 段亮^{1,2}, 王笏辉^{1,2}, 岳昆^{1,2}

(1. 云南大学信息学院, 云南昆明 650500;

2. 云南大学云南省智能系统与计算重点实验室, 云南昆明 650500)

摘要:民间文学文本中含有大量生动形象的修辞手法;人名、地名极其复杂,难以判断词与词之间的边界;与现代汉语表达差别较大,预训练语言模型难以有效地学习其隐含知识,为机器自然语言理解带来困难。该文提出一种基于 BERT 的民间文学文本预训练模型 MythBERT,使用民间文学语料库预训练,将 BERT 的字隐蔽策略改进为对中文词语隐蔽策略。对民间文学文本中解释字、词的注释词语重点隐蔽,减小 BERT 隐蔽的随机性并有利于学习词语语义信息。同时利用注释增强语言模型表示,解决一词多义、古今异义等问题。将 MythBERT 与 BERT、BERT-WWM 和 RoBERTa 等主流中文预训练模型在情感分析、语义相似度、命名实体识别和问答四个自然语言处理任务上进行比较。实验结果表明,注释增强的民间文学预训练模型 MythBERT 在民间文学文本任务上性能显著提升,与基线方法相比取得了最优的效果。

关键词:预训练语言模型;民间文学文本;BERT;自然语言处理;下游任务

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2022)11-0164-07

doi:10.3969/j.issn.1673-629X.2022.11.024

BERT Based Pre-training Model of Folk Literature Texts

TAO Hui-dan^{1,2}, DUAN Liang^{1,2}, WANG Jia-hui^{1,2}, YUE Kun^{1,2}

(1. School of Information Science and Engineering, Yunnan University, Kunming 650500, China;

2. Key Lab of Intelligent Systems and Computing of Yunnan Province, Yunnan University, Kunming 650500, China)

Abstract: There are a large number of vivid figures of speech in folk literature. The named entities, including the name of the people and places, are extremely complicated, which is difficult to judge the boundary between words. It is quite different from modern Chinese expression, which makes it difficult for pre-training language model to learn the implicit knowledge of the texts. It brings great challenges to machine natural language understanding. MythBERT, a pre-training model for folk literature text, is proposed. It not only changes the masked Chinese characters of BERT to the masked Chinese words, but also makes full use of annotations in the folk literature texts to enhance the expression of the language model. MythBERT is compared with some mainstream Chinese pre-training models such as BERT, Bert-WWM and RoBERTa, and four natural language processing tasks were selected for validation, including sentiment analysis, semantic similarity, named entity recognition and question answering. The experimental results show that MythBERT, which enhanced semantics with annotation, has a significant improvement of performance on downstream task for folk literature text, and outperforms the baseline method.

Key words: pre-training language model; Folk literature texts; BERT; nature language processing; downstream task

0 引言

民间文学以“讲唱”形式构成庞大的文本知识体系,融入大量神话、故事与歌谣,是特殊的社会生活方式的汇总。有效提取民间文学信息有助于学者研究民间文学文化,将民间文学文化与商业融合能推动经济发展、激发商业价值。传统民间文学资源与计算机技术有效结合才能凭借新的载体焕发新活力。因此,结

合计算机技术对民间文学文本进行数据整理、挖掘和开发具有重要意义。

预训练模型能够学习文本中的隐含知识并用语言模型进行表示^[1]。大量研究表明,预训练模型有利于提高下游自然语言处理(Natural Language Processing, NLP)任务的性能^[2],对知识图谱^[3]等实际应用有巨大的推动作用。Devlin 等提出预训练模型 BERT^[4](Bi-

收稿日期:2022-06-30

修回日期:2022-08-30

基金项目:云南省重大科技专项(202002AD080002);云南省教育厅科学研究基金(2002Y010);云南大学研究生科研创新项目(2021Y023)

作者简介:陶慧丹(1997-),女,硕士研究生,研究方向为预训练语言模型、机器学习、自然语言处理;通讯作者:段亮(1986-),男,博士,副教授,CCF 会员(95258M),研究方向为社会网络分析、无监督学习、知识发现。

directional Encoder Representations from Transformers) 在 NLP 任务上表现优异。然而,传统预训练模型是由通用领域文本训练而成,无法直接应用于生物医学^[5-7]、金融^[8]和视觉语言^[9]等特定领域文本。此外,BERT 的字隐蔽策略是对输入序列随机隐蔽,民间文学文本中随机隐蔽不能有效地学习到注释词语与句子的关系、建模句子的关键信息和注释句的重要特征。如何利用计算机技术有效地处理民间文学文本,还存在以下挑战:

(1) 特定领域文本与通用领域文本间的巨大差异。民间文学文本语言简洁、表达细腻、内容丰富^[10],会有不断重复语句加强情感表达,而且包含大量专业名词和相关领域的常识性知识。许多词语与现代汉语词语含义相差较大,存在古今异义和一词多义等问题;许多拟人、比喻等修辞手法,蕴含丰富的情感,加大了预训练模型学习民间文学文本深层语义的难度^[11]。

(2) BERT 随机隐蔽策略不适用于民间文学文本。BERT 中所有字的隐蔽概率相同,忽略民间文学文本中注释脚注的重要性。民间文学人名、地名等名词较长,仅对字进行隐蔽会导致词语语义信息的缺失,难以识别词与词的边界。

因此,该文结合 BERT 及民间文学特定领域语料开展预训练,得到民间文学文本的预训练语言模型 MythBERT,主要贡献包括以下几个方面:(1) 利用民间文学文本的注释增强预训练语言模型,将注释脚注中的解释词语替换原句抽象词语,缓解民间文学文本与通用领域文本差异大、一词多义、古今异义、指代关系和隐蔽关系等问题;(2) 利用民间文学文本的注释增强预训练语言模型,考虑全词隐蔽方法(Whole Word Masking),重点关注脚注的注释词语,减小 BERT 隐蔽的随机性,有利于学习词语语义信息;(3) 利用情感分析、语义相似度、命名实体识别和问答四个下游任务对民间文学预训练模型微调,改善实体难以识别边界和修辞手法中复杂情感表达的问题;(4) 使用 BERT 模型的初始权重,减少了重新训练预训练模型带来的巨大开销,并有助于理解民间文学通用知识。在民间文学文本数据集上的实验验证了 MythBERT 的有效性。另外,对不同下游任务进行了测试,进一步证明了 MythBERT 对不同任务的性能都有显著提升。

1 相关工作

预训练模型按照训练文本语料库可以分为通用领域文本和特定领域文本两类。

1.1 通用领域文本预训练模型

通用文本预训练模型使用大规模无标注语料库进行训练以获得文本深层双向语义表示,并通过微调直

接应用于特定 NLP 任务中。BERT 是最具有代表性的预训练模型,大部分模型在此基础上对隐蔽策略、预训练任务、生成任务等进行改进。目前通用文本的预训练模型大部分是 BERT 和基于 BERT 的变种。BERT-WWM(讯飞)^[12]在预训练时使用全词隐蔽策略,以词粒度进行隐蔽;SpanBERT^[13]对随机的相邻分词使用掩码,导致预测掩码困难。RoBERTa(Facebook)^[14]使用精细调参、动态掩码机制等,实验证明 BERT 的下一句子预测任务意义不大;XLNet^[15]使用排序语言模型学习双向上下文语境,解决预训练-微调阶段标记不一致的问题,以大量参数为代价换取效果,提升效果有限;ALBERT^[16]引入句子顺序预测,解决 BERT 的下一句子预测任务低效的问题。ERNIE(1.0)^[17]引入三个阶段屏蔽策略知识,改善了结构化知识问题;ERNIE(THU)^[18]引入知识将实体向量与文本表示融合,但构建知识图谱需要耗费大量的资源。MT-DNN(微软)^[19]利用多个任务微调共享层和任务特定层的参数,但规模巨大、超参数太多不便于调参,需要较多的时间和硬件资源。

双向语言模型使用某种网络作为特征抽取器,将两个不同方向上抽取到的文本表示简单拼接,缺点是只利用了上文或者下文单一的信息,不能同时利用上下文双向信息^[20];隐蔽语言模型作为预训练任务,对堆叠多层的 Transformer 结构难度较低,导致模型无法有效率的学习,并且存在训练阶段有 MASK 标记和微调阶段无 MASK 标记文本不一致的问题,自然语言生成任务中性能较低;排序语言模型保留自回归语言模型的优点,捕获上下文语境,解决训练阶段和微调阶段存在不一致的问题。特定领域的民间文学文本与一般文本在语言表达上存在差异,通用文本预训练模型不能学习专业领域语料库中的术语和表达,无法在特定领域的 NLP 任务中获得高性能。

1.2 特定领域文本预训练模型

在生物医学领域,BioBERT^[5]使用生物医学领域的文章和摘要预训练,评估生物医学任务,有助于其理解复杂的生物医学文献;ClinicalBERT^[6]使用大量临床记录和出院总结文本,提高临床 NLP 任务的性能;SCIBERT^[7]使用大量生物医学领域论文和少量计算机科学领域的论文预训练,评估生物医学 NLP 任务,有助于学习专业领域名词;在金融领域,FinBERT^[8]使用金融新闻和财经文章预训练,增加预训练任务,评估 NLP 任务,捕捉金融领域语言知识和语义信息;在多模态领域,VL-BERT^[9]将视觉和语言作为输入,在大规模的概念标注数据集和纯文本语料库训练,评估视觉 NLP 任务,提高对视觉-语言线索的融合和对齐能力。因此,对特定领域语料预训练,有助于识别特定领

域的专有名词、捕捉常识性知识和语义信息,提高特定领域下游任务的性能。

随机隐蔽会忽略民间文学文本中注释脚注的重要性,对字进行隐蔽导致词语语义信息的缺失。因此,该文采用改进掩码方式的方法,将民间文学注释中的注释释义词语重点隐蔽,通过注释增强语言模型的学习理解能力。传统的语言模型都是基于通用的现代语言语料库无监督训练而来,而民间文学文本中,联合注释加以理解,有助于预训练模型学习更好的语义表示。

2 MythBERT 模型

2.1 模型结构

根据 BERT 输入规则,给定民间文学文本数据集序列 $A = x_1, x_2, \dots, x_m$, 输入序列 $B = y_1, y_2, \dots, y_n$, 增加句首和句子分隔的特殊标记得到 $[\text{CLS}]x_1, x_2, \dots, x_m, [\text{SEP}]y_1, y_2, \dots, y_n, [\text{SEP}]$, $[\text{CLS}]$ 表示句首, $[\text{SEP}]$

表示句子分隔符。面向民间文学文本的 BERT 预训练模型结构如图 1 所示,在 BERT 模型上改进了文本输入预处理方式,对应的 Token Embeddings 也变成 MASK 后的字向量。利用民间文学注释词语(即书籍文本中对难词、难句加以解释的脚注)进行中文分词,对词语 MASK 标记替换。将注释句定位到原文句子尾部,利用注释句对照原句,便于模型加深理解语义。重点关注注释词语,构造预训练任务所需要的训练数据,即 $[\text{MASK}]$ 标记替换得到 $[\text{CLS}]x_1, x_2, \dots, x_m, [\text{SEP}][\text{MASK}], [\text{MASK}], \dots, y_n, [\text{SEP}], [\text{MASK}]$ 表示词隐蔽替换。将输入文本序列中每一个字对应的字向量、分段向量和位置向量相加得到输入向量,输入至多层双向 Transformer 网络,通过自注意力(Self-attention)机制学习文本表示,对其上下文信息进行编码,以预测输入文本中被 MASK 后的词语信息。

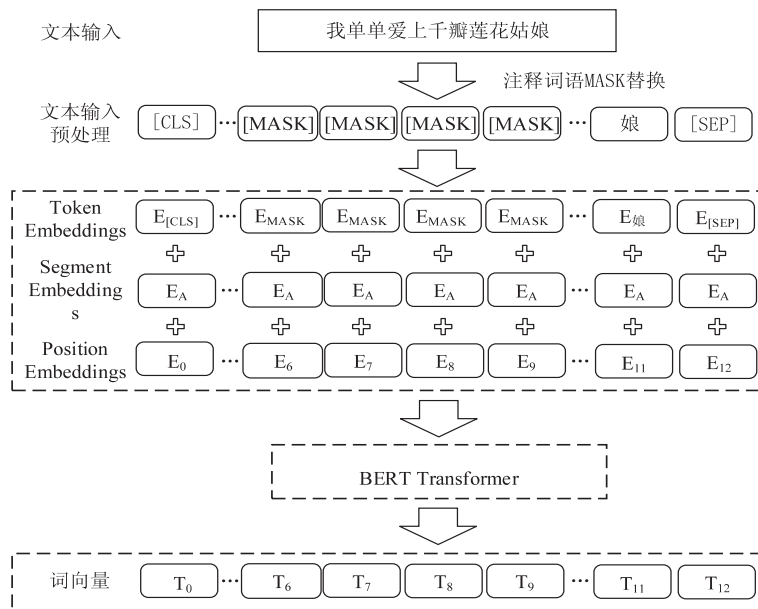


图 1 面向民间文学文本的 BERT 预训练模型结构

2.2 数据预处理

提取民间文学文本注释句中的实体名词便于分词。将注释句分为指代关系注释、古今异义注释和其他注释三种情况,书籍中的注释根据注释序号“①、②、…”依次定位到对应民间文学文本的原句末尾。另外,将指代关系注释和古今异义注释中实体名词替换成注释中的实际意义名词。对预处理后的注释数据添加一个 N 标记,有助于判断是否是注释句,对注释句中的词语进行掩码。

由于民间文学文本中的注释句不多,该文提取注释名词作为字典,对民间文学文本分词,在百度百科和新华词典数据集中搜索分词后有具体意义的实体名词释义。对一词多义的名词释义进行筛选,留下正确的注释并添加到民间文学文本对应句子末尾。

2.3 民间文学文本预训练模型

BERT 的字隐蔽策略是对输入序列随机隐蔽,所有的字隐蔽概率相同。民间文学文本中随机隐蔽不能很好地学习到注释词语与注释句的关系、建模句子的关键信息和学习到注释句的重要特征。因此,MythBERT 对 BERT 隐蔽语言模型的随机隐蔽策略进行改进,对普通词语的隐蔽策略不变,重点关注注释句中的释义词语。

2.3.1 民间文学文本预训练

该文使用官方的 BERT-base(中文)预训练模型的初始权重对民间文学文本语料库预训练。将添加注释处理的民间文学文本数据,经过中文分词后作为数据输入,使用词语隐蔽语言模型,对 BERT 的隐蔽语言模型中的隐蔽策略进行改进,重点关注注释句中的释

义词语。MythBERT 相关符号及含义如表 1 所示。

表 1 符号及含义

符号	含义
A, B	输入序列, B 可为空
T	词向量
$m + n + 3$	序列长度增加句首 [CLS]、句中 [SEP] 和句尾 [SEP] 标记
sentence	序列分词后的数组
n	序列分词后词语的数量
Maxmask	最大 MASK 替换的数量
Notes	注释句的集合, 即民间文学文本中难词、难句的解释。

(1) 普通词语隐蔽策略。MythBERT 和 BERT 使用的隐蔽策略类似, 对输入序列中 15% 的词语替换。其中, 替换的词语有 80% 的概率替换成 [MASK] 标记, 10% 的概率替换成随机词语, 10% 不进行替换。该文对普通词只是将字隐蔽改为词隐蔽, BERT 的随机概率并未改变。

(2) 注释词语隐蔽策略。如果当前处理的句子是民间文学文本原句 (即不含 N 标记), 则对 50% 的概率的注释词语替换成 [MASK] 标记, 另外 50% 不进行替换。

MythBERT 隐蔽策略具体步骤见算法 1。

算法 1: MythBERT 隐蔽策略

输入: $A = x_1, x_2, \dots, x_m, B = y_1, y_2, \dots, y_n$

输出: 词向量 $T = T_0, T_1, \dots, T_{m+n+3}$

步骤:

1. sentence \leftarrow 分词(A, B)
2. FOR $i = 1$ To n DO
3. IF $t > \text{Maxmask}$ THEN
//t 控制序列最大 MASK 的个数
4. break
5. END IF
6. IF $A \in \text{Notes}$ OR $B \in \text{Notes}$ THEN
//注释句则不做 MASK 替换
7. break
8. END IF
9. IF sentence _{i} = 普通词语 THEN
10. 普通词语隐蔽策略
11. END IF
12. IF sentence _{i} = 注释词语 THEN
13. 注释词语隐蔽策略
14. END IF
15. $t \leftarrow t + 1$
16. END FOR

2.3.2 微调 MythBERT

MythBERT 与 BERT 的微调过程相同, 对于每个下游任务, 只需要将各个任务对应的输入和输出送入

MythBERT 结构中。使用民间文学预训练模型只需要将文中模型替换原来的中文 BERT 预训练模型, 不需要更改配置和词汇表文件。该文在以下四个有代表性的民间文学文本挖掘任务上对 MythBERT 进行微调。

(1) 民间文学情感分析对带有强烈情感色彩的文本分析和推理。民间文学文本中带有大量的比喻、拟人的修辞手法, 生动形象地表达主人公的情感色彩。以四句民间文学文本作为一条数据, 分为积极、消极或中性的情感, 标签依次为 1、-1 和 0。

(2) 民间文学语义相似度根据输入的两个句子 A 和 B , 判断其语义是否相似, 意图是否相同。以任意两句作为一条数据, 将文本中的比喻、拟人都判断为语义相同。例如: “可惜我们相差太远” 和 “好像大刀和斧头” 判为语义相同, 标签记为 1, 否则记为 0。

(3) 民间文学命名实体识别。民间文学文本涉及到大量特定领域的专有名词, 识别人、地点、组织是一件非常具有挑战的事情。采用 BIO 标注方法, 将命名实体分为人物 (PER)、地点 (LOC)、组织 (ORG) 和未知实体 (UNK) 四类, 未知实体包括动物、植物、工具等。以句子作为输入, 文本和标签分开存在文件中。

(4) 民间文学问答。从民间文学文本中给出一个问题和一段包含答案的段落, 问答任务输出预测文章答案的跨度。将输入的问题和段落表示为一个单独的序列, 句子 A 表示问题, 句子 B 表示段落。微调时, 起始向量 $S \in R^H$, 结束向量 $E \in R^H$ 。第 i 个单词作为答案跨度开始的概率 P_i 是 T_i 和 S 之间的点积, 然后经过 Softmax 变化得到, 如公式 (1) 所示。

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \quad (1)$$

用户通过问答任务可以根据自己想要了解的民间文学知识进行提问得到解答。

3 实验

该文使用 BERT_{Base} ($L = 12, H = 768, A = 12$) 对民间文学文本进行预训练, L 表示 layers 层数 (即 Transformer 块数), H 表示隐藏层, A 表示自注意力机制的头数。本章将介绍民间文学预训练模型在 4 项 NLP 任务上的测试结果。为了进行公平的预训练模型比较, 每个模型都使用相同的超参数, 预训练时各个模型的初始学习率都设为 $2e-5$, 句子最大长度为 128。微调时初始学习率为 $5e-5$, 最大长度为 128。

3.1 实验设置

(1) 数据集。采用云南大学文学院提供的民间文学文本数据集《云南少数民族古典史诗全集》、《傣族民间故事选》、《娥并与桑洛》和《千瓣莲花》等, 共计

25.3 万条句子。

(2) 测试任务。针对情感分析、语义相似度、命名实体识别和问答任务设置不同评价指标。先进行人工标注,再按照 8:1:1 随机划分训练集、验证集和测试集防止过拟合,具体任务数据集如表 2 所示。评价指标所用到的计算公式如式(2)~式(5)所示。TP 表示正确分类到该类的总数, TP + TN 表示正确分类的总

数, TP + FP 表示预测分类到该类的总数, TP + FN 表示该类的总数。准确率 ACC 表示被预测正确的样本概率;精确率 Precision 表示预测为正确的样本,有多少是真正的正样本。召回率 Recall 表示标记为正的样本,有多少被预测为正。F1 值表示预测答案与真实答案部分一致的匹配程度。

表 2 测试任务数据集

测试任务	任务内容	训练集	验证集	测试集	评价指标
情感分析	分析句子情绪是积极、消极或中性	5.6 k	0.7 k	0.7 k	准确率
语义相似度	判断两个句子的语义是否相同	4 k	0.5 k	0.5 k	准确率
命名实体识别	识别实体,包括人名、地名、机构名和未知实体等	16 k	2 k	2 k	F1 值/精确率
问答	为问题选择对应回答	240		60	F1 值

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

(3) 实验环境。实验基于 Win10 操作系统,采用的 CPU 为 Intel 酷睿 i9-10850K, GPU 为 NVIDIA TITAN V。开发语言 Python3.6,采用深度学习框架 Tensorflow 1.11。

3.2 对比模型

选取中文民间文学数据集,对比 BERT、BERT-WWM、RoBERTa 及该文提出的 MythBERT。

(1) BERT^[4]: 预训练阶段使用隐蔽语言模型和下一句预测任务, MLM 模型对 15% 的 token 进行 mask 标记, 80% 以 [MASK] 标记代替, 10% 以随机 token 代

替以增加噪声, 10% 不改变原始 token。

(2) BERT-WWM^[12]: 如果一个完整词的部分子词被掩码, 则同属完整词的其他子词也会被掩码。

(3) RoBERTa^[14]: 使用精细调参、动态掩码机制等, 将预训练的文本复制 10 份, 每一份随机掩码。同一文本会有 10 种不同的掩码方式, 每个序列被掩码的词不断改变。

3.3 实验结果

将 BERT、BERT-WWM、RoBERTa 和 MythBERT 预训练模型分别在以下四个下游任务上进行了对比, 所有对比模型都在原模型上对民间文学语料库预训练后得到。为了进行公平比较, 对每个数据集, 训练和微调时都使用相同的超参数。该文分别测试超参数 epochs 分别取 2、5、8、10、25、50、100 时对下游任务准确率、精确率和 F1 的影响。四个民间文学自然处理任务结果如表 3 所示。

表 3 四个民间文学自然语言处理任务结果

下游任务	评价指标	BERT		BERT-WWM		RoBERTa		MythBERT	
		DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
情感分析	准确率	82.0	88.0	81.0	88.8	82.0	89.3	83.5	89.8
语义相似度	准确率	78.0	94.8	78.4	95.2	79.6	95.6	82.4	96.8
命名实体识别	F1 值/精确率	66.6/62.2	65.0/60.0	67.3/63.8	65.2/60.7	66.5/62.1	62.0/56.4	68.3/64.7	66.0/61.8
问答	F1 值		31.7		34.2		31.7		36.6

(1) 情感分析。

MythBERT 在情感分析上取得了最好的效果。民间文学验证集准确率达到 83.5%, 对比 BERT、BERT-WWM 和 RoBERTa 分别提升了 1.5 个百分点、2.5 个百分点和 1.5 个百分点; 测试集准确率达到 89.8%, 分别提升了 1.8 个百分点、1.0 个百分点和 0.5 个百分点。因此, MythBERT 有助于捕捉民间文学文本中的内在情感。

不同 epochs 的情感分析准确率如图 2 所示, epochs = 2 时, MythBERT 和 RoBERTa 的准确率比 BERT 和 BERT-WWM 高的多。随着 epochs 的增加, 各个模型的准确率差距缩小, BERT 和 BERT-WWM 收敛较慢。MythBERT 在各个 epochs 的取值时, 情感分析的准确率都是最高的。

(2) 语义相似度。

MythBERT 在语义相似度上取得了最好的效果。

民间文学验证集准确率达到 82.4%, 相较于 BERT、BERT-WWM 和 RoBERTa 分别提升了 4.4 个百分点、4.0 个百分点和 2.8 个百分点; 测试集准确率达到 96.8%, 分别提升了 2.0 个百分点、1.6 个百分点和 1.2 个百分点。因此, MythBERT 有助于学习民间文学文本中的句间关系。

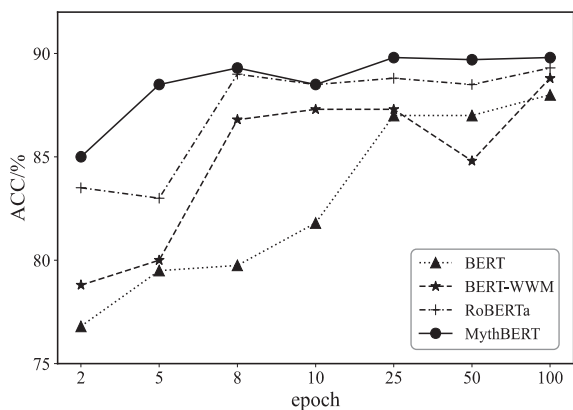


图2 不同 epochs 的情感分析准确率

不同 epochs 的语义相似度准确率如图 3 所示, MythBERT 不断增大 epochs 后, 逐渐稳定在 96.8% 附近。MythBERT 在各个 epochs 的取值时, 语义相似度的准确率都是最高的, 比其他模型更能学习句间关系。

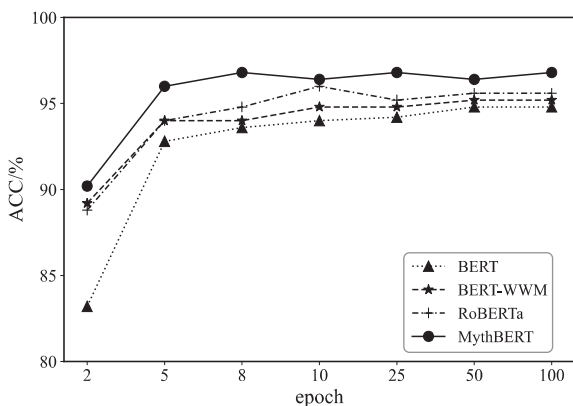


图3 不同 epochs 的语义相似度准确率

(3) 命名实体识别。

MythBERT 在命名实体识别上取得了最好的效果, 有着明显的提升。民间文学验证集 F1 值达到 68.3%, 相较于 BERT、BERT-WWM 和 RoBERTa 分别提升了 1.7 个百分点、1.0 个百分点和 1.8 个百分点; 精确率达到 64.7%, 分别提升了 2.5 个百分点、0.9 个百分点和 2.6 个百分点。测试集 F1 值达到 66.0%, 分别提升了 1.0 个百分点、0.8 个百分点和 4.0 个百分点; 精确率达到 61.8%, 分别提升了 1.8 个百分点、1.1 个百分点和 5.4 个百分点。实验结果表明词 MASK 策略可以更好地识别词与词的边界, 有助于 NER 任务的提升。

不同 epochs 命名实体识别 F1 值如图 4 所示,

MythBERT 的性能在不同 epochs 时均优于其他模型。epochs=5 时, MythBERT 逐渐收敛, 有着上升的趋势。而 RoBERTa 的性能大部分时候远低于其他三个模型。RoBERTa 虽证明去掉 NSP 任务效果更好, 但对代词多, 命名实体复杂和句子关联密切的民间文学文本, NSP 任务至关重要。MythBERT 在预训练时额外增加随机 MASK 注释词语, 因此, 模型训练收敛, 需要更多的 epochs。

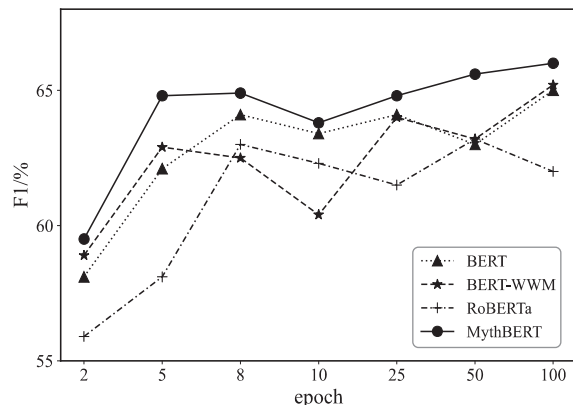


图4 不同 epochs 命名实体识别 F1 值

(4) 问答。

民间文学测试集 F1 值达到 36.6%, 比 BERT、BERT-WWM 和 RoBERTa 分别提升了 4.9 个百分点、2.4 个百分点和 4.9 个百分点, MythBERT 与其他模型相比有着显著的提升。推测与命名实体识别任务的提高有关, 答案大多来源于实体名词, 且词语隐藏策略都比 BERT 有着明显的提升效果。

不同 epochs 的问答 F1 值如图 5 所示, 当 epochs 为 2 时, MythBERT 性能大大领先于其他模型。随着 epochs 不断增大, 其他模型也相继收敛, 逐渐逼近 MythBERT。但 MythBERT 在各个 epochs 值都仍然优于其他模型。

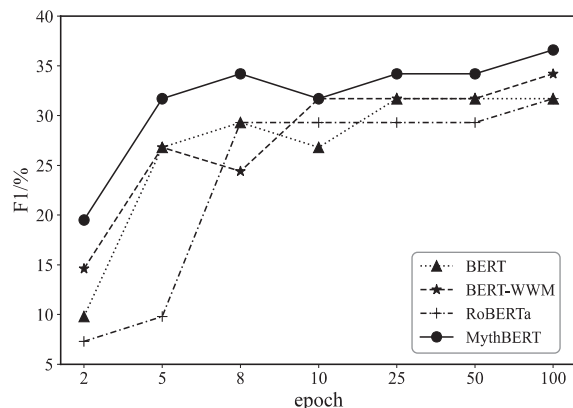


图5 不同 epochs 的问答 F1 值

4 结束语

该文提出了一种基于注释增强的民间文学文本预

训练模型 MythBERT, 该模型改进了 BERT 的隐蔽语言模型策略, 对民间文本中的注释词语进行重点关注, 并通过情感分析、语义相似度、命名实体识别和问答这四个下游任务对民间文学预训练模型微调。在上述四个任务上的实验验证了 MythBERT 的有效性, 尤其是在命名实体识别和问答任务上有较大提升。提出的方法能以较低成本构建民间文学领域的预训练模型, 该思路也可应用到那些具有较多注释的文本中, 如文言文书籍等。该文的下游任务还集中在民间文学数据集, 在数据集规模、预训练语言模型对比、下游任务对比、模型性能评价指标等各个方面还有待拓展。

参考文献:

- [1] 王乃钰, 叶育鑫, 刘露, 等. 基于深度学习的语言模型研究进展[J]. 软件学报, 2021, 32(4): 1082–1115.
- [2] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: a survey[J]. Science China: Technological Sciences, 2020, 63(10): 1872–1897.
- [3] 王鑫, 邹磊, 王朝坤, 等. 知识图谱数据管理研究综述[J]. 软件学报, 2019, 30(7): 2139–2174.
- [4] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis: [s. n.], 2019: 4171–4186.
- [5] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234–1240.
- [6] ALSENTZER E, MURPHY J R, BOAG W, et al. Publicly available clinical BERT embeddings[J]. arXiv: 1904.03323, 2019.
- [7] BELTAGY I, LO K, COHAN A. SciBERT: a pretrained language model for scientific text[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing. Hong Kong, China: [s. n.], 2019: 3613–3618.
- [8] LIU Z, HUANG D, HUANG K Y, et al. FinBERT: a pre-trained financial language representation model for financial text mining[C]//Proceedings of the twenty-ninth international joint conference on artificial intelligence. Yokohama: [s. n.], 2020: 4513–4519.
- [9] SU W J, ZHU X Z, CAO Y, et al. VL-BERT: pre-training of generic visual-linguistic representations[J]. arXiv: 1908.08530, 2019.
- [10] 吴斌, 吉佳, 孟琳, 等. 基于迁移学习的唐诗宋词情感分析[J]. 电子学报, 2016, 44(11): 2780–2787.
- [11] 宋挺, 郭展成, 何世柱, 等. 基于动态词遮掩的句子匹配预训练模型[J]. 中文信息学报, 2021, 35(11): 43–50.
- [12] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. arXiv: 1906.08101, 2019.
- [13] JOSHI M, CHEN D, LIU Y, et al. SpanBERT: improving pre-training by representing and predicting Spans[J]. arXiv: 1907.10529, 2019.
- [14] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized BERT pretraining approach[J]. arXiv: 1907.11692, 2019.
- [15] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Proceedings of advances in neural information processing systems 32: annual conference on neural information processing systems. BC: [s. n.], 2019: 5754–5764.
- [16] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[C]//Proceedings of 8th international conference on learning representations. Addis Ababa: [s. n.], 2020.
- [17] SUN Y, WANG S, LI Y, et al. ERNIE: enhanced representation through knowledge integration[J]. arXiv: 1904.09223, 2019.
- [18] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities[C]//Proceedings of the 57th conference of the association for computational linguistics. Florence: ACL, 2019: 1441–1451.
- [19] LIU X, HE P, CHEN W, et al. Multi-task deep neural networks for natural language understanding[C]//Proceedings of the 57th conference of the association for computational linguistics. Florence: ACL, 2019: 4487–4496.
- [20] 岳增营, 叶霞, 刘睿珩. 基于语言模型的预训练技术研究综述[J]. 中文信息学报, 2021, 35(9): 15–29.