文章编号:

结合预训练模型和语言知识库的文本匹配方法

周烨恒 石嘉晗 徐睿峰*

（哈尔滨工业大学（深圳） 计算机科学与技术学院，广东省 深圳市 518055）

摘要：针对文本相似性匹配任务，该文提出了一种大规模预训练模型融合外部知识库的方法。该方法分为三阶段：基础语言模型预训练阶段、外部知识库学习任务生成及联合训练阶段、下游任务微调阶段。在该文中探讨了方法的设计原理和原则。面向第二阶段，利用 WordNet 语言知识库生成学习任务，提升了现有 BERT 模型的性能。在第三阶段，针对相似度匹配任务进行微调。该文还试验了对知识库生成的学习任务和引入的外部任务进行联合训练，通过在微软公司提出的 MT-DNN 模型基础上取得进一步的性能提升，证明了知识库学习任务联合强化的可行性。此外，该文还探讨和验证了生成学习任务时结合下游任务任务特定知识，以在下游任务无法微调时改进模型性能的方法。

关键词：文本匹配；预训练模型；语言知识库融合

中图分类号：TP391 **文献标识码：**A

Text matching method combining pre-trained model and language knowledge base

Yeheng Zhou , Jiahan Shi , Ruifeng Xu*

(School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen , Guangdong 518055, China)

Abstract: For the text similarity matching task, this paper proposes a method for large-scale pre-trained model to integrate the external language knowledge base. The method can be divided into three stages: the primary language model pre-training stage, the external knowledge base learning task generation and joint training stage, and the downstream task fine-tuning stage. The design principles of the method are discussed in this paper. For the second phase, the WordNet language knowledge base was used to generate learning tasks, which improved the performance of the existing BERT model. In the third phase, fine-tuning was performed for the similarity matching tasks. This paper also tested the joint training of the learning tasks generated by the knowledge base and the imported external tasks. Further performance improvement based on the MT-DNN model proposed by Microsoft Corporation, proved the feasibility of the joint reinforcement of knowledge base learning tasks. Besides, this paper explored and validated methods for generating learning tasks in conjunction with downstream task-specific knowledge to improve model performance when downstream tasks cannot be fine-tuned.

Key words: text matching; pre-trained model; language knowledge base fusion

0 引言

在自然语言处理过程中，经常会涉及到如何对文本之间的相似性进行匹配的需求。文本的相似性度量在许多领域有着广泛的应用，包括信息检索，文本分类，阅读理解，机器问答乃至深层

收稿日期：2019-06-11；**定稿日期**：2019-08-12

基金项目：国家自然科学基金（U1636103;61632011;61876053）；深圳市基础研究项目（JCYJ20180507183527919，JCYJ20180507183608379）；深圳市技术攻关项目（JSGG20170817140856618）；深圳证券信息联合研究计划资助；哈尔滨工业大学（深圳）创新研修课资助

通讯作者：徐睿峰

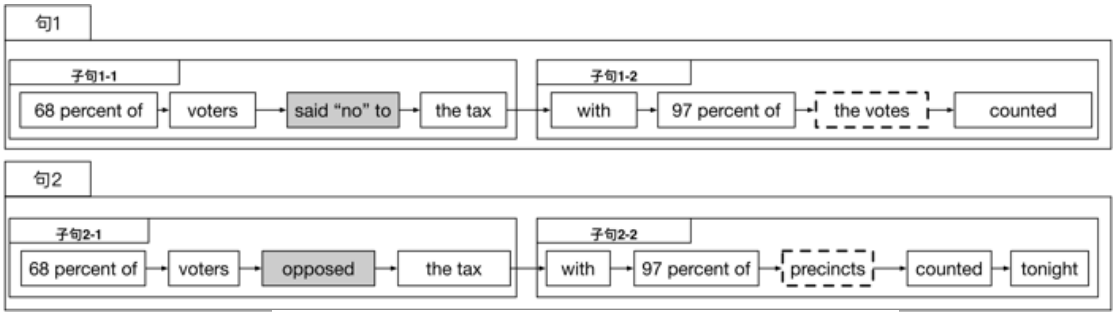

图 1 现有预训练模型相似匹配遗漏识别示例

次的语义理解等。

以 Google 的 BERT[1]、OpenAI 的 GPT2[2] 等为代表的大规模预训练语言模型，能够通过第一阶段预先在巨型规模语料上多层联合学习上下文隐藏信息，和第二阶段后续针对具体下游任务做微调的两阶段方法，利用捕捉到的丰富语言信息，提升多种自然语言处理任务的效果，且大幅度刷新了通用语言理解评估（GLUE）基准评测[3]的各项指标。特别是，当预训练模型作用于相似度匹配任务时，其性能达到了新的高度。

尽管这些大规模预训练模型，已经帮助各类任务取得了惊人的提升，也成为很多任务、新方法事实性的基础组件，但是它们在语言理解上仍然存在诸多盲区：包括对语言知识、领域知识和常识知识理解能力的欠缺[4]。

首先，在语言知识方面，我们发现相似度匹配结果显示，模型常常会因词汇、词组级别语言知识缺失产生错漏。如图 1 所示，句 1 和句 2 上各自由两个子句组成，虽然对应子句之间仅在少部分实体和触发词之间存在表示的区别：如子句 1-1 和子句 2-1 之间存在触发词“said ‘no’ to”和“opposed”的差异；子句 1-2 和子句 2-2 之间存在实体“the votes”和“precincts”之间的差异。实际上，以人类的视角，从语义层次上容易判别它们是相似的。然而，现有的预训练模型，一方面没有对这类内在联系做针对性训练，另一方面尽管预训练语料规模很大，但是蕴含该层次内部联系的内容却并不足够多，模型在预训练中有限相关内容里也难以学会它们，最终缺少相关语义知识。这样，当进入评测阶段，很容易仅仅因为个别词汇表示不同，在文本相似度匹配任务中判定句 1 和句 2 不相似。

其次，尽管各项评测常常假设文本匹配任务存在部分标注语料，能够进行有监督的微调，可是在各项实际应用中，该假设通常会因受到诸如：边缘计算设备性能限制、高质量语料缺乏、任务的高响应速度要求等，多方面现实需求和限制的组合约束而不成立。

我们注意到，当该假设不能满足时，预训练模型由于缺乏相似度匹配的领域常识知识、不理解什么是相似度匹配、如何进行相似度匹配、匹配指标应该如何评估，其性能会出现较大的倒退，甚至根本无法有效进行该类任务。

对于模型中特定层次知识缺失，尤其是相似度匹配任务中预训练模型在同义词组级别语义知识缺失的问题，比较直截了当的方案是进一步扩充预训练语料集，虽然该方法可以实现间接扩充包含该类型知识的语言内部表示的语料，以促使模型逐渐学习相关知识（如图 2，其可行性也被 GPT-2 证实[2]）。然而，在现有大规模训练模型训练的时间成本、经济成本已然极高的基础上，如果再扩大语料，进一步提高的成本将达到一个惊人的新高度，一般的研究机构和企业根本无力承担。

另一种巧妙的方案是，在现有大规模预训练模型基础上通过设计任务训练来融入外部知识库知识。这一思路的动机是，联想到人类学习过程，其大多是在先通过早期的学习形成一个初步的语言模型基础，再逐渐通过针对性的任务训练，来学习某些领域的知识和任务技能，最终形成在相关领域的世界观和方法论，实现对相关问题的解决。具体而言，在相似度匹配任务的语言知识缺失和匹配方法知识缺失上，解决方案可以是利用该层级外部语言知识库，设计包含相似度匹配训练的任务以实现融入缺失知识。一方面，

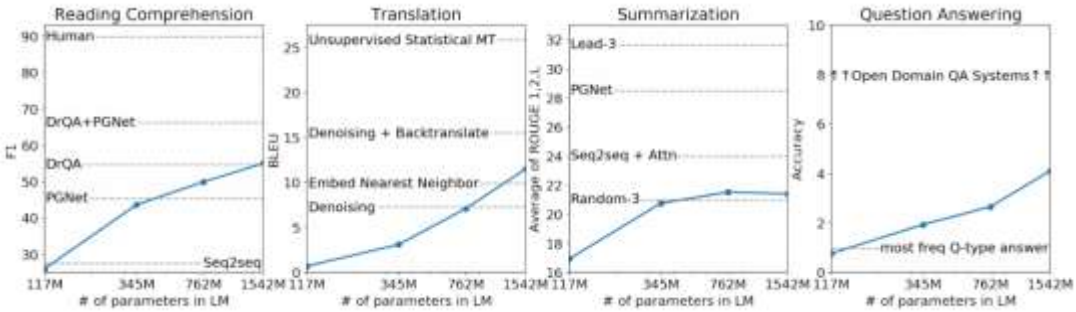

图 2 部分 NLP 任务微调性能与预训练语料规模关系[2]

前人在语言知识库方面已经做了大量的工作，如 WordNet[5]、HowNet[6]等，有了相对丰富的积累；另一方面，相对众包标注等渠道质量参差不齐的数据[7]，外部语言知识库本身源自于专家研究，也经过广泛和长期的实践检验，知识库质量很高。

相对于前一种方案，方案二虽然在知识库选择和任务设计等环节需要一定的人工参与，但是可以让成本大幅度降低，能够面向具体知识缺失针对性提升性能。除此之外，如果设计的任务本身如果能够包含可以指导如何进行具体任务方法（如相似度匹配任务方法）的知识，那么在不存在标注语料，或其他有监督微调假设不能成立的场合，将会帮助性能实现极大的提升。

在本文中，为了解决相似度匹配任务中的这些问题，我们将利用方案二，提出一种基于基础语言模型预训练、外部知识库联合强化、下游任务微调三阶段的知识库语义强化方法。实验结果表明，本方法对相似度匹配任务具有强化能力。

1 相关工作

对于预训练模型，学术界已有很多的研究，主要分为特征和微调两大类别，试图从训练语料中获取潜在的语言模型信息。

早期的工作以词向量为代表，包括 Collobert 和 Weston[8]、Mikolov 等人[9]、Pennington 等人[10]，聚焦在基于特征的方法上。主要思路是通过捕获词语的某些特征，将词语转换成向量空间中的离散表示，然后用作各种模型的嵌入。由于典型语言模型中的词汇存在上下文相关性，Peters 等人的 ELMo[11]采用两个不同方向的序列模型结合来捕获上下文相关的复杂特征。这些方法仅仅是将语言模型的集成作为特征简单的引入任务模型中。

而自 Dai 和 Le 将无标签语料上的预训练得到的框架和参数作为下游任务开始点起[12]，越来越多的研究关注到基于微调的方法上。Radford 等人[13]提出了利用一个单项生成预训练的 Transformer[14]来学习语言表示。Devlin 等人基于自注意的多头多层双向 Transformer，并结合超大规模数据集，提出了 BERT[1]，在多项自然语言处理任务上都取得了 state-of-art 的成果。而 OpenAI 的 GPT-2 模型则是将下游任务从有监督的微调改为无监督、结合继续增加层数和语料的方法，在生成相关任务取得巨大改进[2]。微软公司的工作则证明了多任务联合方法可以适用于微调阶段，并带来性能的累进增强[15]。

尽管这些预训练模型在很多领域取得重大成功，但是在语言知识、领域知识和通用知识的语义理解和利用上还存在很多的挑战和缺失。近期的一些研究以及开始关注相关问题：Baidu 提出的 ERNIE[15] 模型将尝试百科、文库作为通用外部实体知识库语料大规模引入模型。Huang[16]和 Beltagy[17]分别尝试引入临床医学和科学知识，将领域知识引入模型。

对于文本相似度和匹配相关的研究，早期主要集中在通过基于分类体系和统计特征的两种类别方法上。基于分类体系的方法主要是通过以树、图为主的特殊结构，结合语义词典和世界知识，语义词典和世界知识本身也常被组织成树状层次结构[18]，这些结构中的路径特征可以作为某种语义的度量。Lodhi、Saunders 等人[19]的提出了序列核方法，Wang、Ming 等人[20]利用语法树

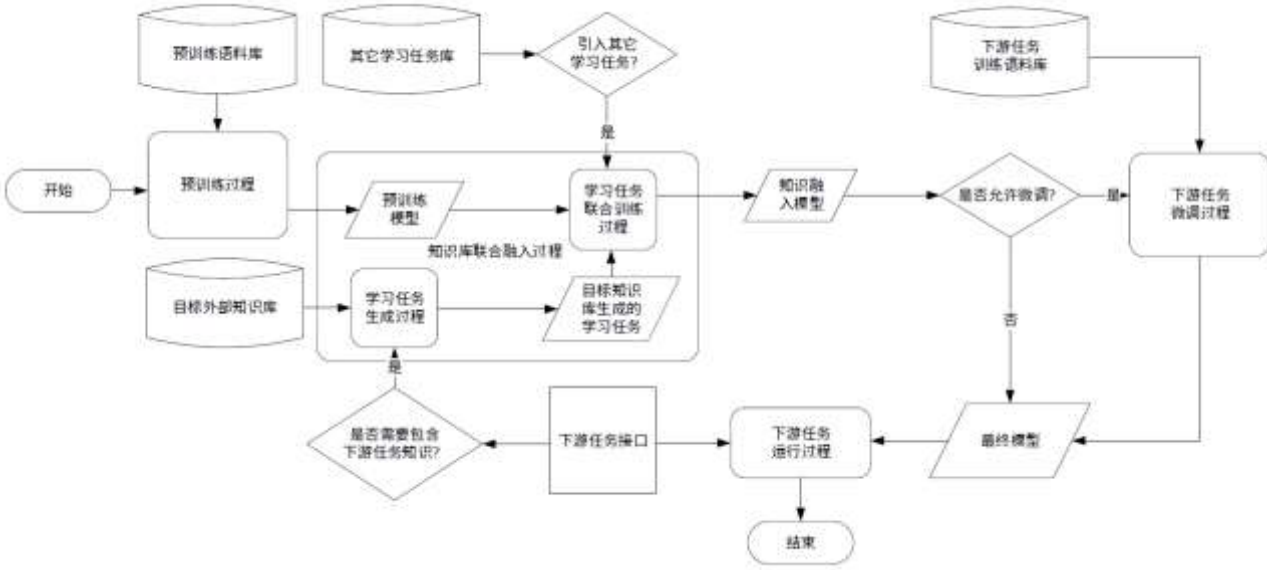

图 3 三阶段知识库融入方法整体流程的示意图

匹配方法寻找相似问题，Culotta 和 Sorensen 等人提出一种核计算两颗依存树相似性[21]。而统计方法方面，Salton 和 Buckley 提出 TD-IDF，其词语重要性随文件内频率提升，随语料库库出现频率反比下降的方法[22]启发了大量利用语言学统计特征进行文本相似度匹配的研究，如 J Mueller 和 A thyagarajan 将 RNN 网络引入句子相似性任务中[23]。Moreas、Baki 等将多种树核作为 SVM 特征利用联合方法计算相似性解决文献信息提取问题[24]。预训练模型出现后，这些模型本身也将文本相似性匹配任务的指标大幅度刷新[1][2][15]，Hu 等人研究了数种预训练模型作用于文本匹配任务的效果[25]，Wataru Sakata 等人将 BERT 用于常见问题解答（FAQ）中查询问题和答案的相似性[26]。

预训练模型应用于文本相似度任务，本质上是预训练捕获了大量的隐式语言联系和部分背景知识，它同时包含语言模型的统计特征，结合了基于分类体系和统计特征方法的长处。但是现阶段其针对性获取语义词典和背景知识的能力仍然较弱，现有研究融入这些知识也仅仅限于少数领域知识或者通用命名实体知识的层面。而本文尝试从语言知识库生成文本相似度和掩膜预测任务的形式，同时学习语言知识和相似度匹配任务知识。此外，受 MT-DNN[15]启发，我们将学习知识库的多个任务联合以期取得累加的效果。

2 方法

本节中，我们将展示：整个方法的三阶段流程；多语言知识库学习任务联合训练方法；利用 WordNet 同义、反义关系生成相似度匹配学习任务的方法；利用 WordNet 词组、固定用法生成掩膜预测学习任务的方法。

2.1 整体流程

该方法的流程图如图 3 所示，总体上分为三个阶段：预训练阶段，知识库融入阶段，下游任务微调阶段。

其中第一阶段使用现有预处理模型，其主要追求目标为在量大、质量优良、种类丰富的巨型语料上学习基础语言模型、少量基本常识，并形成语料积累。

第二阶段可以分为两个环节：①首先由专家针对需要融入的知识库及其知识的特征设计生成知识学习任务。需要注意的是，视下游任务需求，如果需要，此时可以考虑在生成的学习任务中包含下游任务。②针对生成的学习任务和用以辅助学习的其它任务（可选）进行联合训练。

第三阶段则是针对下游具体任务进行微调，在本文中为相似度匹配任务。如果需要模拟性能限制、缺乏下游训练集等不能进行微调的场景，该阶段可省略，直接输出第二阶段输出的结果。

2.2 多知识库学习任务联合训练

算法 1：多知识库学习任务联合训练

随机初始化模型参数为θ
读取预训练模型
设定最大世代数为$\text{epoch}_{\max}$
for t in 1,2,..,T do
    打包数据集dataset_t到微型批中D_t中
end
for e in 1,2,..,epoch_{max} do
    合并所有数据集
    $D = D_1 \cup D_2 \ldots \cup D_T$
    随机打乱 D
    for b_t in D do
        计算损失：$L(\theta)$
        $L(\theta)$= 任务损失
        计算梯度：$\nabla(\theta)$
        更新模型：$\theta = \theta - \epsilon\nabla(\theta)$
    end
end

在通过第二阶段第一环节后，利用目标知识库针对性生成了一个或者多个学习任务，还可能引入了其它任务库中选出的学习任务。对于多个学习任务，此时如果依次训练，可能出现灾难性的遗忘，且如果将联合学习任务训练视为下游的联合微调的等价，则已经联合训练已经被证明优于单方向训练[15]，故此处选用联合训练方法。其具体算法见算法 1，此处 T 为为生成的和引入的学习任务，dataset_i 为 T_i 任务所对应的训练数据集（和任务一起生成）。

2.3 同反义知识-相似匹配学习任务

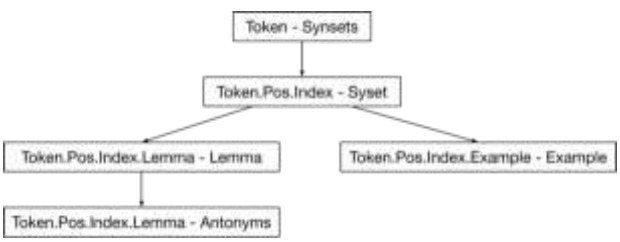

图 4 WordNet 同义反义词结构示意

如图 4 所示，WordNet 的每个词（Token）可以拥有自己的同义集合集（Synsets），集合中的每个同义集合（Synset）表示为 Token.Pos.Index，Token 为对应的具体语法形式，Pos 为词性，包括名词、动词、形容词、副词四大类，Index 表示是该集合该词性中第几种语义，每个集合中包含词义对（Lemma）表示一组义和形的唯一组合。另外，每一个 Lemma 都可能含有数量不等的反义词，每一个 Synset 也关联了一组用例集合。

如算法 2，扫描 WordNet 词表，对词表中的每个词获取同义集合集。然后遍历同义集合集中集合，首先得到该词集词性，接着得到该词集关联用例，再得到所有词义对。对关联用例集去除非该词用例，以防止同义集用例重复；对词义对集去除非同词性词和词集表示自身，去重并防止后续语法变形推导错误。

算法 2：基于 WordNet 的反义同义语料生成

$G = \{\}$
    for w in Words:
        S_w = 获取同义集的集合(w)
        for s in S_w:
            pos_s = 获取词性(s)
            E_s = 获取用例集(s)
            L_s = 获取词义对集(s)
            F_s = 标记词集变形规则集(s,pos_s)
            for e in E_s:
                $GS_s = \{\}$
                $GA_s = \{\}$
                f_{es} = 解析变形规则(e, s,F_s)
                for l in L_s:
                    gs_l = 变形生成例句(l,e,r_{es})
                    GS_s += gs_l
                    A_l = 获取反义集(l)
                    for a in A_l:
                        ga_a = 变形生成例句(a,e,r_{es})
                        GA_s += ga_a
    $G = G \cup \{(0,gs,ga) \mid \text{for gs,ga in } (GS_s \cdot GA_s)\}$
    $G = G \cup \{(1,g1,g2) \mid \text{for g1,g2 in } (GS_s \cdot GS_s)\}$

对每个同义词集的表示，生成符合其词性所

有可能的变形。遍历过滤后的关联用例集，对每个用例搜索所有符合该词变形的子串，标记变形属性（词性、时态、人称、词位等）、起止位置。嵌套遍历过滤后的同义词义对集，对每个同义词对，按照标记的变形属性和起止位置进行变形和替换，加入同义生成组；再嵌套遍历每个同义词对的反义词集，按照标记的变形属性和起止位置进行变形和替换，加入反义生成组。

将同义生成组合反义生成组组合，同义组成员之间标注为句对相似匹配标签“1”，同-反义组成员之间标注为标签“0”，且由于同义生成组的数量一般远远大于反义生成组，故在生成标签“1”时，将同义词集表示本身和组内其它所有生成语句匹配，而同义生成组合反义生成组之间完全两两匹配，以减少最终生成的数据集合标签的不均衡程度。

值得注意的是，上述按照标记变形过程中，由于 WordNet 中存在大量的词组，不能简单运用现有库以单词的形式直接变形，而需要针对不同词性特性额外处理。

2.4 词组-变长掩膜预测学习任务

和真实世界中的语言对应，WordNet 中同样存在大量的词组/固定用法。如图 5，现存主要预训练模型，如 BERT 是以 Token 为输入和掩膜单位，每次遮蔽 15%的词来预测以学习语言模型。然而这样就可能丢失某些固定词组的结构特征和隐式语义，或者是需要大大增加捕捉该组合的信息所需的语料量和计算代价。

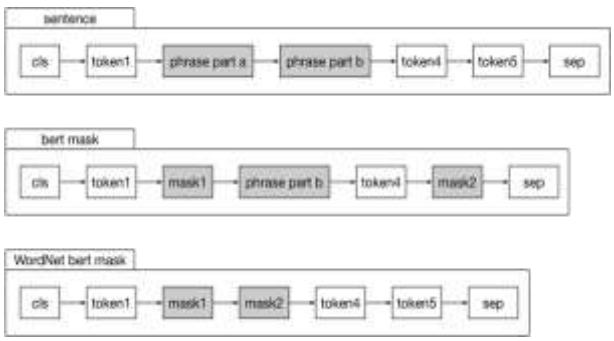

图 5 BERT 字掩模和基于 WordNet 词组掩模

为解决这一问题，可以采取以词组/固定用法为输入和如图 5 中 WordNet-Based Mask 以词组/固定用法为掩膜的两种解决手段。但是，前者又面临着在要修改模型结构和分词级联误差的问题，那么就剩下变长掩膜的办法，Google 的 n-gram 掩膜 BERT 也是相似的思路。然而，由于 WordNet 本身已经确定性地指出词组/固定用法，那么自然可以确定性得到该区段其掩膜长度，又可以减少 n-gram 滑动不同尺寸窗口的计算量。

需要指出的是，同样由于 WordNet 中有些用例较短，或者不是完整句子，可以通过搜索引擎和额外词典数据库的方式，以这些用例为种子，获取完整用例乃至扩充词组/固定组合学习语料集。

3 实验

在本节中，我们将会详细描述实验的设置并分析实验结果。由于预训练成本较高，且现有预训练模型和本方法第一阶段兼容，本实验的预训练模型采用已存在的预训练语言模型 BERT-Base；对于学习任务联合强化实验，使用的 MT-DNN 同样基于 BERT-Base。

3.1 实验数据集和超参数设置

实验数据集采用微软发布的 Microsoft Research Paraphrase Corpus（MRPC）[27] 和 Quora 发布的 Quora Question Pairs（QQP）[28]数据集。这两个数据集都是面向文本对、判断文本信息等价性的英语数据集。

其中，为保持简明和高效，所有实验都将至少包含在较小的 MRPC 数据集上测试的结果。对于部分提升不是特别显著的实验，为增强结论的说服力，在规模更大的 QQP 上数据集上也进行测试。

具体任务类型、数据集划分和评测指标见表 1：

表 1 实验数据集设置

数据集	任务	训练集	测试集	指标
MRPC	Paraphrase	4K	1.7K	F1
QQP	Paraphrase	404K	391K	F1

3.2 节中无训练集的任务知识学习能力实

验，BERT + WordNet 强化学习任务，迭代次数 1，知识库生成的相似性匹配任务最大序列长度 64，掩膜预测任务最大序列长度 128，batch 大小 4。BETR 微调中微调，迭代次数 1，任务最大序列长度 64，batch 大小 4。3.3 节中为，其它参数不变，只有微调集合的迭代次数改变为 10。

3.2 无训练集的任务知识学习能力实验

在引言中，我们根据经验判断当下的预训练模型由于没有针对性的训练特定下游任务，其预训练语料中含有隐式反映下游任务做法的语料也不够充足，当出于终端设备性能、缺乏标注数据等各种现实限制，必须无微调地面对新的下游任务时，它们将不“知晓”如何进行这些下游任务判断和决策的知识，因而相对于进行微调的有监督，效果很差。我们还在方法流程设计中，作出了根据下游任务形式设计知识库学习任务，不仅能学习到知识库，而且能够学习到下游任务的“方法论”知识，最终提升下游任务无法进行监督微调情况下的性能表现的猜想。

为了验证这一对假设和猜想的正确性，针对 MRPC 数据集设计了如下实验。首先只使用 BERT 预训练模型，直接对下游的文本相似匹配任务测试集进行测试，和在预训练模型上使用下游任务训练集进行有监督微调后再测试的结果进行对比；接着尝试只在 BERT 预训练模型上进行第二阶段的 WordNet 相似度匹配学习任务和词组/固定搭配掩膜预测学习任务的联合学习，而下游不进行有监督微调，只进行测试。

表 2 无训练集的任务知识学习能力实验结果

方法	MRPC
BERT 无微调	0
BERT 无微调 + WordNet	**0.778**
BERT	0.786

上述结果中，BERT 预训练模型在不经过下游任务训练集的有监督微调时，在 MRPC 相似性匹配任务中将全部结果全部分到‘0’（不相似）一组，而 BERT 经过训练集微调后即可迅速提升性能，说明此时模型尚未“学会”如何进行文本相似匹配任务，证明猜想一。

而在 BERT 预训练模型基础上，使用利用 WordNet 生成的学习任务（包含相似性匹配）学习，而不利用训练集进行微调的实验组也能在测试集上取得远高于仅使用 BERT 预训练，而近似于 BERT 微调的成绩的成绩，一定程度上证明了根据下游任务形式设计知识库学习任务可以帮助“学习下游任务特定知识”， 提升下游任务无法提供有监督微调情况下的性能表现。

3.3 预训练模型的知识库强化实验

本节实验要证明在下游任务有微调的情况下，知识库生成的学习任务增强的预训练模型能够进一步提升在文本匹配任务上的表现，分别在 MRPC、QQP 两个数据集上进行了实验。

性能基准主要参照人类表现[3]和 XLNET-Large(ensemble)[29]模型，后者为当前最好模型，相较于 BERT-Base 乃至 BERT-Large 都大大扩大了预训练的语料规模，改进了模型结构，并使用了集成方法。此外，还比较了以下传统文本匹配方法：CBOW，使用 Glove[30]的词袋集成；Skip-Throught，使用 TBC 训练用于预测前一句和下一句的序列到序列模型编码器[31]；BiLSTM + CoVe + Attn，将注意力机制结合到双向长短时记忆网络（BiLSTM）、上下文向量[32]；InferSent，使用 MNLI、SNLI 训练最大池化的 BiLSTM[33]；BiLSTM + ELMO + Attn，基于注意力机制使用 ELMo 表示[11]的 BiLSTM。

表 3 预训练模型知识库强化实验

方法	MRPC	QQP
CBOW	0.734	0.791
Skip-Throught	0.717	0.822
BiLSTM + CoVe + Attn	0.718	0.834
InferSent	0.741	0.817
BiLSTM + ELMO + Attn	0.780	0.843
Bert	0.843	0.892
BERT + WordNet	**0.849**	**0.895**
XLNet-Large (ensemble)	0.907	0.902
Human Performance	0.808	0.804

结果显示，基于 BERT 预训练模型，同样在整个训练集上进行微调，使用 WordNet 生成的学

习任务学习知识库知识进行强化的方法能够进一步在 MRPC 和 QQP 数据集上相较于基准的 BERT 微调分别提升 0.6%和 0.3%的模型性能。不仅超过全部传统方法和人类表现，还在不大幅扩大预训练语料和不使用集成方法的情况下，进一步缩小了与当前最佳模型 XLNET-Large（ensemble）的性能差距。

出于展现小数据量上增强能力的考虑，另外针对 MRPC 数据集设置了一组仅选用训练集前 100 条数据进行微调的对照实验，实验结果如表 4：

表 4 预训练模型知识库强化实验（小训练集）

方法	MRPC（训练集前 100 条微调）
BERT	0.784
BERT + WordNet	**0.796**

表 4 结果显示，在 MRPC 数据集上，将微调的范围限制在训练集前一百条数据后，知识库强化较 BERT 微调的性能提升为 1.2%，比在整个数据集上微调的同比提升更加明显，可能是训练集本身也包含了知识，扩大训练集后知识库增强的边际效用减弱了。

接下来，为具体研究各知识库学习任务的效果，在 MRPC 数据集上对设计的两个知识库学习任务，即文本匹配学习任务和掩模预测学习任务进行消融实验，实验结果如表 5：

表 5 WordNet 知识库学习任务的消融实验

方法	MRPC
BERT	0.843
BERT + WordNet 文本匹配学习任务	0.844
BERT + WordNet 掩模预测学习任务	0.847
BERT + 文本匹配/掩模预测生成任务	**0.849**

如表 5 显示，单独使用 WordNet 掩模预测学习任务能够在 MRPC 数据集 BERT 微调基础上将性能提升 0.4%；而单独使用 WordNet 文本匹配学习任务提升较小，观察测试集分类结果，发现尽管对同反义表示误分类情况有所缓解，但是处理子句结构等价变形的能力有所下降，说明同反义文本匹配学习任务的生成算法设计仍有改进空间。最后，综合使用两种生成的学习任务，取得了累进的提升效果。

3.4 知识库学习任务联合强化实验

3.2 节中已经证明了根据下游任务特殊设计的知识库学习任务能够帮助模型获得该类任务的任务特定知识。那么从这种意义上，各种下游任务都可以视为某种形式的知识库生成任务。MT-DNN 证明了下游任务的联合训练可以实现性能的累进提升[15]，那么同样地可以假设知识库生成任务和引入的外部任务的联合训练，也能够达到类似的效果。实验 3.4，通过在第二阶段 WordNet 生成的两种任务的基础上，加入 MT-DNN 中的其它任务，并进行联合训练，在 MRPC、QQP 数据集上测试其性能。性能基准参照为 BERT 微调方法、人类表现[3]和 XLNET-Large(ensemble)[28]模型。

表 6 知识库学习任务联合强化实验

方法	MRPC	QQP
Bert	0.843	0.892
MT-DNN	0.861	0.896
MT-DNN + WordNet	**0.865**	**0.899**
XLNet-Large (ensemble)	0.907	0.902
Human Performance	0.808	0.804

实验结果显示 WordNet 知识库的融入能在 MT-DNN 的多任务联合基础上，进一步提升 MRPC、QQP 数据集上文本匹配任务的性能。证明上述猜想的合理性，也说明多个知识库生成任务和引入的其它任务的联合训练的设计有其合理性。

4 结论与展望

在本文中，我们使用 BERT 等现有预训练模型，针对文本相似性匹配任务存在的语言知识缺失及在无微调训练集下存在的任务特定知识缺失问题，提出了一种预训练语言模型、知识库生成学习任务增强、下游任务微调的三阶段方法：说明了第二阶段多个目标知识库生成任务和外部任务联合训练的方法；解释了使用 WordNet 知识库同义、反义知识生成相似度匹配学习任务，词组知识生成掩膜预测学习任务的具体方法；通过无训练集情况的预训练模型和知识库生成任

务增强后，在测试集上的性能对比，说明该方法可以帮助学习任务特定知识；通过在 MRPC、QQP 数据集上，对 BERT 和 MT-DNN 模型强化后进一步提升的性能说明了知识库可帮助强化，且联合强化方案是合理的。

既然 WordNet 知识库生成学习任务，联合训练后可以让 BERT 和 MT-DNN 预训练模型，实现对下游的文本匹配任务的强化，那么，该方法或许可以拓展到其它预训练模型、其它知识库、其它下游任务上去。在未来的研究中，可以进一步探究该方法迁移的可行性，乃至能否成为一个通用方法框架。

参考文献

[1] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv:1810.04805.

[2] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[EB/OL].https://www.techbooky.com/wp-content/uploads/2019/02/Better-Language-Models-and-Their-Implications.pdf.

[3] Wang A, Singh A, Michael J, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding[J]. arXiv preprint,2018, arXiv:1804.07461.

[4] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced Language Representation with Informative Entities[J].arXiv preprint, 2019, arXiv:1905.07129.

[5] Miller G A. WordNet: An electronic lexical database[M]. MIT press, 1998.

[6] 董强，董振东.《知网》.[DB/OL].http://www.keenage.com.

[7] Hsueh P Y, Melville P, Sindhwani V. Data quality from crowdsourcing: a study of annotation selection criteria[C]//Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. Association for Computational Linguistics, 2009: 27-35.

[8] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th International Conference on Machine Learning. ACM, 2008: 160-167.

[9] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural information Processing Systems. 2013: 3111-3119.

[10] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[11] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint, 2018, arXiv:1802.05365.

[12] Dai A M, Le Q V. Semi-supervised sequence learning[C]//Advances in Neural Information Processing Systems. 2015: 3079-3087.

[13] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding with unsupervised learning[R]. OpenAI, 2018.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[15] Liu X, He P, Chen W, et al. Multi-Task Deep Neural Networks for Natural Language Understanding[J].arXiv preprint, 2019, arXiv:1901.11504.

[16] PaddlePaddle, ERNIE[CP/OL]. https://github.com/PaddlePaddle/LARK/tree/develop/ERNIE.

[17] Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission[J]. arXiv preprint, 2019, arXiv:1904.05342.

[18] Beltagy I, Cohan A, Lo K. SciBERT: Pretrained Contextualized Embeddings for Scientific Text[J]. arXiv preprint, 2019, arXiv:1903.10676.

[19] Lodhi H, Saunders C, Shawe-Taylor J, et al. Text classification using string kernels[J]. Journal of Machine Learning Research, 2002,

2(Feb): 419-444.

[20] Wang K, Ming Z, Chua T S. A syntactic tree matching approach to finding similar questions in community-based qa services[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009: 187-194.

[21] Culotta A, Sorensen J. Dependency tree kernels for relation extraction[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 423.

[22] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information rocessing & management, 1988, 24(5): 513-523.

[23] Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings of 30th AAAI Conference on Artificial Intelligence. 2016.

[24] Moraes L, Baki S, Verma R, et al. University of Houston at CL-SciSumm 2016: SVMs with tree kernels and Sentence Similarity[C]//Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). 2016: 113-121.

[25] Hu W, Dang A, Tan Y. A Survey of State-of-the-Art Short Text Matching Algorithms[C]//Proceedings of International Conference on Data Mining and Big Data. Springer, Singapore, 2019: 211-219.

[26] Sakata W, Shibata T, Tanaka R, et al. FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance[J]. arXiv preprint, 2019, arXiv:1905.02851.

[27] Mircosoft, Microsoft Research Paraphrase Corpus[DB/OL].https://www.microsoft.com/en-us/download/details.aspx?id=52398.

[28] Quora. Quora question pairs[DB/OL]. https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs.

[29] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding[J]. arXiv preprint, 2019, arXiv:1906.08237.

[30] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[31] Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors[C]//Proceedings of Advances in Neural Information Processing Systems. 2015: 3294-3302.

[32] McCann B, Bradbury J, Xiong C, et al. Learned in translation: Contextualized word vectors[C]//Proceedings of Advances in Neural Information Processing Systems. 2017: 6294-6305.

[33] Conneau A, Kiela D, Schwenk H, et al. Supervised learning of universal sentence representations from natural language inference data [J]. arXiv preprint, 2017, arXiv:1705.02364.

周烨恒（1998-），本科生，主要研究领域为自然语言处理、语义计算和知识图谱。

Email：master@evernightfireworks.com

石嘉晗（1997-），本科生，主要研究领域为自然语言处理、文本情感分析。

Email：shijiahan@stu.hit.edu.cn

通讯作者：

徐睿峰（1973-），博士，教授，博士生导师，主要研究领域为自然语言处理、文本情绪计算、认知计算。

Email：xuruifeng@hit.edu.cn