文章编号：1003-0077(2020)02-0063-10

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 深圳 518055）

摘　要：针对文本匹配任务，该文提出一种大规模预训练模型融合外部语言知识库的方法。该方法在大规模预训练模型的基础上，通过生成基于 WordNet 的同义—反义词汇知识学习任务和词组—搭配知识学习任务引入外部语言学知识。进而，与 MT-DNN 多任务学习模型进行联合训练，以进一步提高模型性能。最后利用文本匹配标注数据进行微调。在 MRPC 和 QQP 两个公开数据集的实验结果显示，该方法可以在大规模预训练模型和微调的框架基础上，通过引入外部语言知识进行联合训练有效提升文本匹配性能。

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

A Text Matching Method by Combining Pre-trained Model and Language Knowledge Base

ZHOU Yeheng, SHI Jiahan, XU Ruifeng

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

Abstract: Aiming at text matching task, this paper proposes a method to incorporate large-scale pre-training model and external language knowledge base. On the basis of large-scale pre-training model, this method introduces external linguistic knowledge by generating synonym-antonym knowledge learning task and phrase-collocation knowledge learning task based on WordNet, respectively. Then, the two new generated tasks are joint trained with MT-DNN multi task learning model to further improve the model performance. Finally, the annotated text matching data is used to fine tune. The experimental results on two open datasets, MRPC and QQP, show that the proposed method can effectively improve the performance of text matching by introducing external language knowledge for joint training on the basis of the framework of large-scale pre-training model and fine-tuning.

Keywords: text matching; pre-training model; language knowledge base

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、机器问答乃至深层次的语义理解等。

2018 年以来，以 Google 的 BERT[1]、OpenAI 的 GPT2[2]等大规模预训练语言模型为第一阶段，以针对具体下游任务进行微调为第二阶段的框架，取得了很好的性能。这一框架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，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文本匹配任务时，其性能达到了新的高度。

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

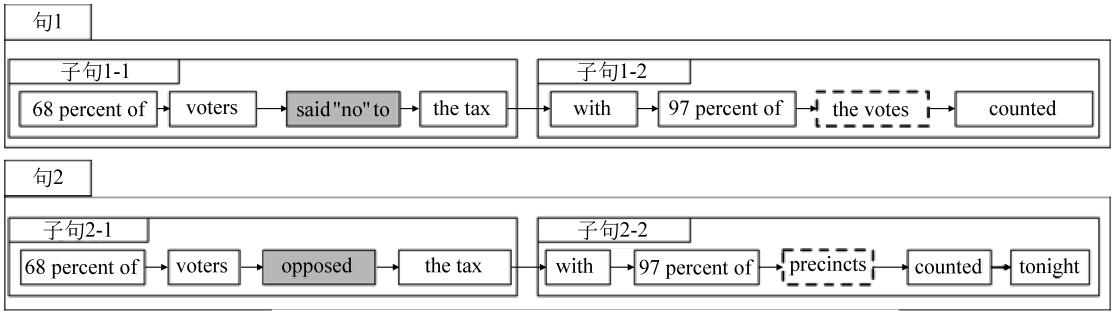


图 1 现有预训练模型的相似文本匹配遗漏识别示例

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的欠缺[4]。

首先，在语言知识方面，我们发现文本匹配结果显示，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。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，很容易仅仅因为个别词汇表示不同，导致在文本匹配任务中判定句 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，基于预训练模型的框架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，尤其是文本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，以促进模型逐渐学习到相关知识。然而，在现有大规模预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，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。

另一种方案是，在现有大规模预训练模型框架基础上融入外部语言知识库。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基础的语言模型，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，最终获得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，在文本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，设计包含文本匹配训练的任务以实现融入缺失知识。一方面，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，第二种方案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，但是可以让成本降低，还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包含指导如何进行具体任务方法（如文本匹配任务方法）的知识，那么在缺乏标注语料的情况下将会帮助性能实现提升。

基于第二种方案，本文提出一种基于基础语言模型预训练、外部知识库联合学习、下游任务微调三阶段的方法。主要通过生成基于 WordNet[5] 的知识学习生成任务引入语言学知识，并与外部任务进行联合训练。实验结果显示，本方法有效利用了外部语言知识，提高了文本匹配性能。

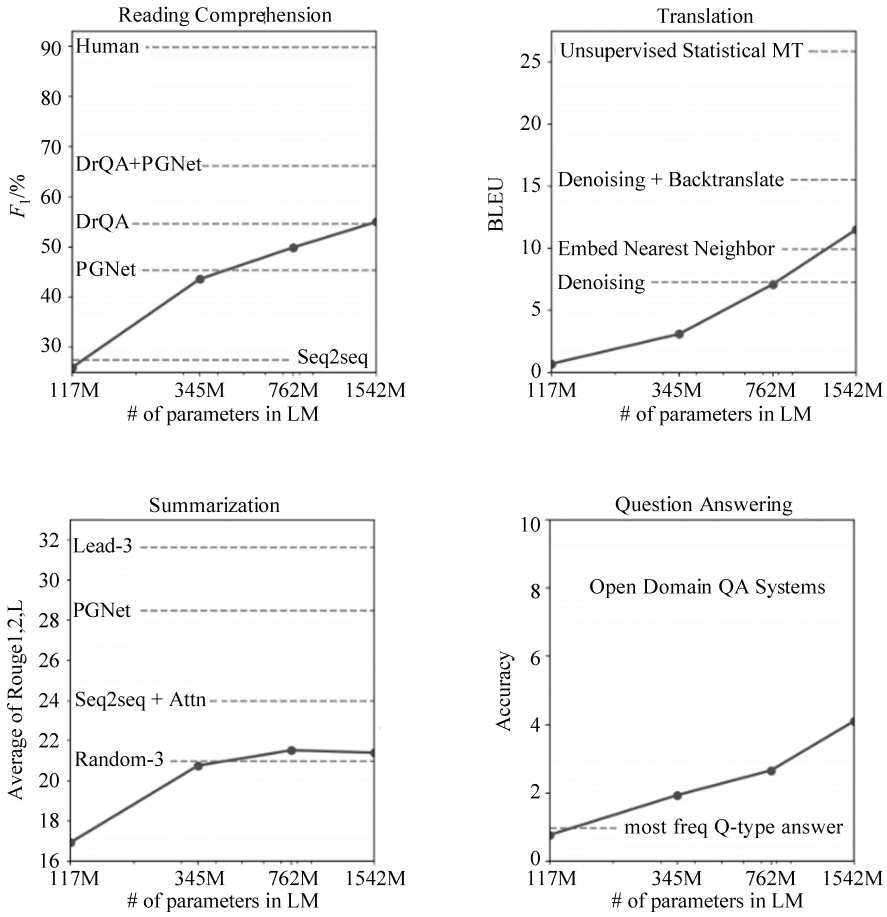


图 2　部分 NLP 任务微调性能与预训练语料规模关系[2]

1　相关工作

对于预训练模型，学术界已有很多的研究。早期的工作以词向量为代表，包括 Collobert 和 Weston[8]、Mikolov 等[9]、Pennington 等[10]的工作，主要聚焦在基于特征的方法上。其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地引入任务模型中。

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[13]提出了利用一个单向生成预训练的 Transformer[14]来学习语言表示。Devlin 等人基于自注意力的多头多层双向 Transformer，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-the-art 的性能。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，并结合继续增加层数和语料的方法，在文本生成相关任务取得巨大改进[2]。微软公司的 MT-DNN 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。近期的一些研究已经开始关注相关问题：Baidu 的 ERNIE[16]初步尝试将旗下百科、文库作为通用外部实体知识库语料大规模引入模型。Huang 等[17]和 Beltagy 等[18]分别尝试引入临床医学和科学知识，将领域知识引入模型。

文本匹配相关的研究。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。由于语义词典和世界知识本身可以被组织成树状层次结构[5]，因此这些结

构中的路径特征可以作为某种语义的度量。Lodhi、Saunders 等[19]提出了序列核方法；Wang 等[20]利用语法树匹配方法寻找相似问题；Culotta 和 Sorensen 提出一种核函数计算两颗依存树相似性[21]。在基于统计特征的方法方面，Salton 和 Buckley 提出 TD-IDF[22]。J Mueller 和 A Thyagarajan 将 RNN 网络引入文本匹配任务中[23]；Moraes 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]。预训练模型的出现，将文本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]。

预训练模型能够应用于文本匹配任务，本质上是由于预训练捕获了大量的隐式语言联系和部分世界知识。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。但是其获取语义词典和世界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。

2 方法

本节中，我们将说明：整个方法的三阶段流程；多个知识库学习任务联合训练的方法；利用 WordNet 生成同义—反义词汇知识和词组—固定搭配知识学习任务的方法。

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。

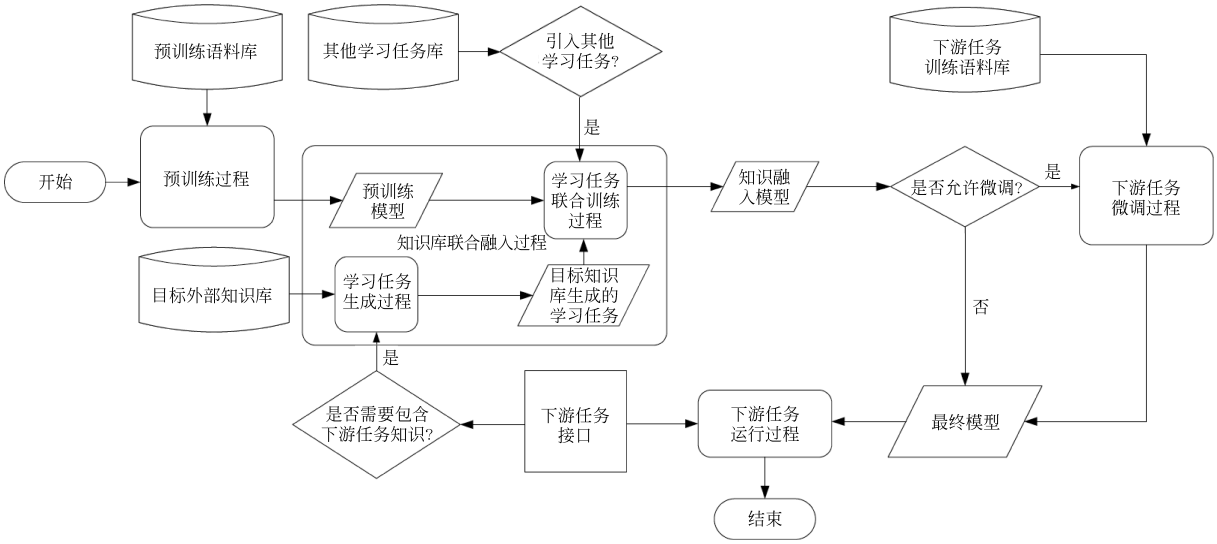


图 3 三阶段知识库融入方法整体流程的示意图

第一阶段使用现有预处理模型，其主要追求目标为在规模大、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首先根据需要融入的语言知识库及其知识的特征生成知识学习任务，这里分别基于 WordNet 生成同义—反义词汇知识的学习任务和词组—固定搭配知识学习任务。②针对生成的学习任务和可选引入的、用以辅助学习的其他外部任务(可选)进行联合训练。

第三阶段，如果能够获得面向具体的下游文本匹配任务的训练集，则针对该训练集进行微调。如果遇到终端性能限制或者缺乏下游训练集的场景，则可以跳过微调阶段，直接利用第二阶段训练得到的模型进行具体的下游文本匹配任务。由于第一阶段和第三阶段目前较为成熟，在此不再赘述，重点介绍本文核心的第二阶段方法。

2.2 多知识库学习任务联合训练

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。如果将联合学习任务训练视为下游的联合微调的等价，则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 为生成和引入的学习任务，$dataset_t$ 为和任务 t 一起生成的对应训练数据集。

算法 1：WordNet 知识学习与外部引入任务联合训练

输入：*model* 预训练模型
输入：$epoch_{max}$ 最大世代数
输入：*tasks* 多知识库学习任务输出
输出：联合训练后的模型
function Training（*model*，$epoch_{max}$，*tasks*）
θ←Get Reference of model params
T← Size of *Tasks*
for t←1 to T **do**
　$dataset_t$←Get the datasets of $task_t$
　D_t←Initialize mini-batch
　D_t←Pack $dataset_t$ into mini-batch
end for
for *epoch* ← 1 to $epoch_{max}$ **do**
　$D \leftarrow D_1 \cup D_2 \cup \cdots \cup D_T$
　D←Shuffle D
　for t←1 to T **do**
　　b_t←Get a mini-batch of $task_t$
　　$L(\theta)$←Compute loss of θ
　　$\nabla(\theta)$ ←Compute gradient with $L(\theta)$
　　$\theta \leftarrow \theta - \nabla(\theta)$
　end for
end for
return *model*
end function

为了进一步提升任务性能，这里我们还选择 MT-DNN 选用的 GLUE 基准性能评估[3]中的其他任务进行多任务联合训练，以进一步提高文本匹配性能。需要注意的是，因需要评估文本匹配任务性能，这里需要排除 MT-DNN 中包含的文本匹配任务。

2.3　同义—反义词汇知识的学习任务生成

针对预训练模型在同义—反义词汇知识方面的不足，如果能有效利用 WordNet 知识库以同义词集的方式组织的结构，充分发掘其表达反义关联的关系，将可以弥补该类知识缺失，有效提升性能。此外，WordNet 中的同义—反义词汇知识，天然地蕴含着一部分文本相似性、尤其是句对相似性的度量，能够帮助模型针对性学习这种文本匹配任务的特定知识。

在 WordNet 中，每个词（Token）可以拥有自己的多个同义集合（Synsets）。每个同义集合（Synset）以一个合适的概念词（Token.Pos.Index）表达该同义集合的概念。其中 Token 为对应的具体词形，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。

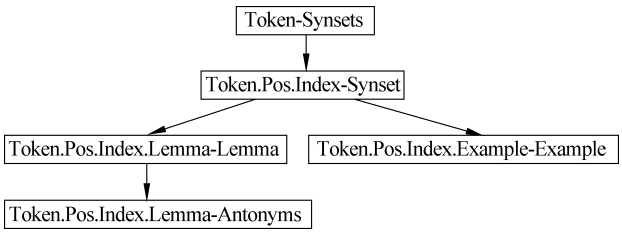


图 4　WordNet 同义—反义词结构示意

在一个同义集合中包含若干个词义对（Lemma），表示一组义和形的唯一组合。另外，每一个词义对都可能含有数量不等的反义词，构成个反义集合（Antonyms）。每一个同义集合也关联了一组用例，构成用例集合（Examples）。

算法 2 扫描 WordNet 词表 *tokens*，对词表中的每个词 t 获取其关联的全部同义词集 *ss*。然后对每个关联的同义词集 s，分别取该同义词集词性 *pos*，该词集关联用例 *es*，该词集所有词义对 ls。其中，对关联用例集去除非该词用例，使得每个用例只被使用一次，防止重复；去除与表示同义词集的概念词词性不同的词义对，防止后续产生语法变形的推导错误；去除表示同义词集的概念词本身，防止生成前后完全相同的句对。

对表示同义词集的概念词，生成符合其词性所有可能的变形规则 *rs*。对过滤后的关联用例集的每个用例 e，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，得到变形规则 r。其中，对过滤后的每个词义对 l，按照标记的变形属性和起止位置进行变形和替换，生成新同义用例，加入同义用例组 *synes*。对每个词义对的反义词集 *as* 中的词 a，按照标记的变形属性和起止位置进行变形和替换，生成新的反义用例，加入反义用例组 *antes*。

将同义用例组 *synes* 的同义用例成员 *se*、*se*1 两两组合，构造同义句对（*se*，*se*1，1），其匹配标签“1”，加入生成任务语料 *corpus* 中。将同义用例组 *synes* 成员与反义用例组 *antes* 的成员 *se*、*ae* 两两组合，构造反义句对（*se*，*ae*，0），其匹配标签为“0”，加入生成任务语料 *corpus* 中。由于同一个同义词集的同义用例集的成员数量，一般远远大于其反义用例集的成员数量，故在生成拥有标签“1”的同义句对时，仅仅将表示同义词集的概念词所生成的同义用例和其它词汇生成的同义用例匹配。而在生成拥有标签“0”的反义句对时，完全的两两匹配，以减少最终生成的同义—反义句对的不均衡程度。

算法 2 同义—反义词汇知识学习任务语料生成

输入：*WordNet* 知识库
输出：*corpus* 生成的学习任务语料

```
function GenTaskCorpus(WordNet)
  corpus←ø
  tokens←Get tokens of WordNet
  for i←1 to Size(tokens) do
    t←tokens[i]
    ss←Get the synsets of token
    for j←1 to Size(ss) do
      s←ss[j]
      pos←Get the pos of s
      es←Get the examples of s
      ls←Get the lemmas of s
      rs←Get transforming rules of s and pos
      for k←1 to Size(es) do
        e←es[k]
        synes←ø
        antes←ø
        r←Parse transforming rules of e, s, rs
        for x←1 to Size(ls) do
          l←ls[x]
          synes←synes∪Generate l,e,r examples
          as←Get the antonymous of l
          for y←1 to Size(as) do
            a←as[y]
            antes←antes∪Generate a,e,r examples
          end for
        end for
        for x←1 to Size(synes) do
          se←synes[x]s
          for y←1 to Size(synes) do
            se1←synes[y]
            corpus←corpus∪(se,se1,1)
          end for
          for z←1 to Size(antes) do
            ae←antes[z]
            corpus←corpus∪(se, ae,0)
          end for
        end for
      end for
    end for
  end for
end for
return corpus
end function
```

续表

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额外处理。

2.4 词组—固定搭配知识学习任务生成

与同义—反义词汇知识缺失类似，预训练模型常常错误区分词组—固定搭配边界，导致文本匹配任务的性能下降。究其原因，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。

为解决这一问题，可以采取以词组/固定用法为输入和以词组/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，可以减少 n-gram 滑动不同尺寸窗口的计算量。

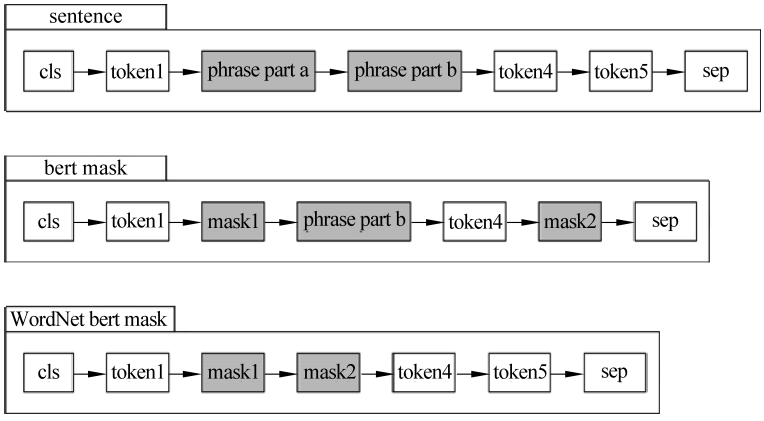


图 5 BERT 掩模和 BERT WordNet 掩模示例

需要指出的是，由于 WordNet 中有些用例较短、或者不是完整句子，需要通过搜索引擎和额外词典数据库的方式，以这些用例和种子，获取完整用例乃至扩充词组—固定搭配组合学习语料。

3　实验

3.1　实验设置

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判断文本对信息等价性的文本匹配数据集。具体任务类型、数据集划分和评测指标见表 1。

表 1　实验设置

数据集	任务	训练集	测试集	指标
MRPC	Paraphrase	4 K	1.7 K	F_1
QQP	Paraphrase	404 K	391 K	F_1

本实验的预训练模型采用预训练语言模型 BERT-Base。无训练数据的文本匹配实验中，在 WordNet 生成学习任务和对照微调的迭代次数都设为 1，batch 大小为 4。其中，WordNet 生成的同义—反义词汇知识学习任务最大序列长度 64；词组—固定搭配知识学习任务最大序列长度 128，batch 大小 4，对照的微调最大序列长度为 64。在结合训练数据进行微调的实验中，其他参数不变，只有微调集合的迭代次数统一改变为 10。

3.2　无训练数据的文本匹配实验

面向无下游任务训练数据的情况，针对 MRPC 数据集设计了三组实验。第一组只使用 BERT 预训练模型，直接应用下游的文本匹配任务进行评估；第二组在预训练模型基础上，利用 WordNet 生成的两个学习任务进行学习后，直接进行文本匹配评估；第三组是对照组，在预训练模型基础上，使用文本匹配任务的训练数据进行有监督微调，并进行评估。实验结果如表 2 所示。

表 2　无训练数据的文本匹配实验结果

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

上述结果中，第一组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为‘0’（不相似）一组。而第三组经过训练集微调后即可迅速提升性能，说明预训练模型本身尚未“学会”如何进行文本匹配任务。

而第二组在 BERT 预训练模型基础上，使用利用 WordNet 生成的学习任务进行学习，而不利用训练集进行微调。其成绩远高于第一组，仅次于有训练数据微调的第三组，一定程度上证明了根据下游学习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”，提升下游任务的性能表现。

3.3　有训练数据的文本匹配实验

本组实验评估在有训练语料情况下，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。

性能基准指标如下：

CBOW[29]　使用 GloVe 的词袋集成；

Skip-Throught[30]　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；

BiLSTM＋CoVe＋Attn[31]　将注意力机制结合到双向长短时记忆网络（BiLSTM）、上下文向量；

InferSent[32]　使用 MNLI、SNLI 训练最大池化的 BiLSTM；

BiLSTM＋ELMO＋Attn[11]　基于注意力机制使用 ELMo 表示的 BiLSTM；

GLUE Human Baselines[3]　GLUE 排行榜上提供的人工基准；

XLNet-Large（ensemble）[33]　当前最好模型，扩大了预训练的语料规模，改进了模型结构，因其使用了多种模型集成方法，在这里仅作为参考。

表 3　有训练数据的文本匹配实验

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**
HumanPerformance	0.808	0.804

实验结果如表3所示，其中预训练模型全部在下游任务数据集上进行有监督微调。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所示。

表4 有训练数据的文本匹配实验(小训练集)

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

表4显示，在MRPC数据集上，将微调的范围限制在训练集前100条数据后，知识库强化较BERT微调的性能提升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，即同义—反义词汇知识和词组—固定搭配知识学习任务进行消融实验，实验结果如表5所示。

表5 WordNet知识库学习任务的消融实验

方法	MRPC
BERT	0.843
BERT+WordNet同义—反义词汇知识学习任务	0.844
BERT+WordNet词组—固定搭配知识学习任务	*0.847*
BERT+WordNet生成的两个学习任务	**0.849**

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节的无训练数据的文本匹配实验已经表明，根据下游任务设计的知识学习任务能够帮助模型获得任务特定知识。从这种意义上，各种下游任务都可以视为某种形式的知识学习任务。通过在WordNet生成的两种任务的基础上，加入MT-DNN中非文本匹配的其他任务[6]，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、GLUE Human Baselines[3]和XLNet-Large(ensemble)[33]模型。实验结果如表6所示。

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**
HumanPerformance	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。多个知识库生成任务和引入的其他任务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，针对文本匹配任务存在的语言知识缺失和任务特定知识缺失问题，提出了一种预训练语言模型、知识学习任务增强、下游任务微调的三阶段方法。

面向文本匹配任务，分别在有训练数据和无训练数据的情况下，通过设计基于WordNet的同义—反义词汇知识学习任务和词组—固定搭配知识学习任务，以及引入外部知识进行多任务联合训练的方法，有效提升了在MRPC、QQP数据集上的性能表现。

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他预训练模型、其他知识库、其他下游任务上去。在未来的研究中，可以进一步探究该方法迁

移的可行性。

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,arXiv: 1810.04805,2018.

[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, arXiv: 1804.07461,2018.

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

[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]//Proceedings of 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, arXiv: 1802.05365,2018.

[12] Dai A M,Le Q V. Semi-supervised sequence learning [C]//Proceedings of 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]//Proceedings of 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,arXiv: 1901.11504,2019.

[16] Paddle, 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,arXiv: 1904.05342,2019.

[18] Beltagy I,Cohan A,Lo K. SciBERT: Pretrained contextualized embeddings for scientific Text[J]. arXiv preprint,arXiv: 1903.10676,2019.

[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 Processing & 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, arXiv: 1905.02851,2019.

[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] 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.

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

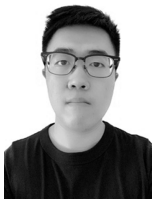
[31] McCann B, Bradbury J, Xiong C, et al. Learned in translation: Contextualized word vectors[C]//Proceedings of Advances in Neural Information Processing Systems. 2017: 6294-6305.

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

[33] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized autoregressive pretraining for language understanding [J]. arXiv preprint, arXiv: 1906.08237,2019.

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。
E-mail: master@evernightfireworks.com

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。
E-mail: xuruifeng@hit.edu.cn



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。
E-mail: shijiahan@stu.hit.edu.cn