北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
ISSN 0479-8023,CN 11-2442/N

《北京大学学报(自然科学版)》网络首发论文

题目：复述平行语料构建及其应用方法研究
作者：王雅松，刘明童，张玉洁，徐金安，陈钰枫
DOI：10.13209/j.0479-8023.2020.078
收稿日期：2020-06-07
网络首发日期：2020-10-15
引用格式：王雅松，刘明童，张玉洁，徐金安，陈钰枫．复述平行语料构建及其应用方法研究．北京大学学报(自然科学版).
https://doi.org/10.13209/j.0479-8023.2020.078

北京大学学报(自然科学版)
Acta Scientiarμm Naturaliμm Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2020.078

复述平行语料构建及其应用方法研究

王雅松　刘明童　张玉洁†　徐金安　陈钰枫

北京交通大学计算机与信息技术学院，北京 100044；† 通信作者，E-mail: yjzhang@bjtu.edu.cn

摘要　复述指同一语言内相同意思的多样化表达。自然语言理解的难点在于语言表达的多样性，机器难以正确理解和处理所有的表达。复述能够涵盖更多的语言现象和多样化的语义表达，对增强信息检索、机器翻译、自动问答等自然语言处理任务的性能有十分重要的作用。构建多样性的复述数据，对复述相关研究以及在自然语言处理任务中的应用具有十分重要的意义。因此，本文以汉语为研究对象，提出构建大规模高质量汉语复述平行语料的方法，基于翻译引擎进行复述数据增强，将英语复述平行语料迁移到汉语，同时人工构建汉语复述评测数据集。基于构建的汉语复述数据，我们在复述识别和自然语言推理任务上验证复述数据构建及其应用方法的有效性。首先基于复述语料生成复述识别数据集，预训练基于注意力机制的神经网络句子匹配模型，训练模型捕获复述信息。然后，将预训练的模型用于自然语言推理任务，改进其性能。我们在自然语言推理公开数据集上进行评测，实验结果表明，本文构建的复述语料可以有效应用在复述识别任务中，模型可以学习复述知识。当应用在自然语言推理任务中，复述知识能够有效地提升自然语言推理模型的精度，验证了复述知识对下游语义理解任务的有效性。我们提出的复述语料构建方法，不依赖具体语言，可以为其他语言和领域提供更多的训练数据，生成高质量的复述数据，以进一步改进其他任务的性能。

关键词　复述语料构建；数据增强；迁移学习；复述识别；自然语言推理

Research on the Construction and Application of Paraphrase Parallel Corpus

WANG Yasong，LIU Mingtong，ZHANG Yujie†，XU Jin'an，CHEN Yufeng

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;
† Corresponding author, E-mail: yjzhang@bjtu.edu.cn

Abstract　Paraphrases refer to different expressions of the same meaning in the same language. The difficulty of natural language comprehension lies in the diversity of languages. It is difficult for machines to correctly understand and process different expressions of the same semantics. Paraphrase cover more linguistic phenomena and diverse semantic expressions, which plays a very important role in improving the research of natural language processing tasks such as information retrieval, machine translation, and automatic question answering. To increase the diversity of paraphrase data, and verify that paraphrase in the role of natural language processing tasks, with Chinese as the research object, put forward the method to construct large-scale and high-quality paraphrase parallel corpora. The method include transfering English paraphrase corpus to Chinese, by using the method of translation engines, and manually annotating test set to build the evaluation data set. Based on the constructed Chinese paraphrase data, we verify the validity of the paraphrase and construction method in the paraphrase recognition task and natural language inference task. Firstly, the paraphrase recognition data is generated based on the constructed paraphrase corpus, and the

国家自然科学基金(61876198，61976015，61976016)资助
收稿日期: 2020–06–07; 修回日期: 2020–08–15

attention-based neural network model of sentence matching is pre-trained to capture the paraphrase information. Then, the pre-trained model is applied to the natural language inference task. We measure the open data set. The experimental results show that the constructed paraphrase corpus can be effectively applied to the paraphrase recognition task, and the model can learn paraphrase knowledge. When applied to natural language inference task, paraphrase knowledge can effectively improve the accuracy of natural language inference models and verify the effectiveness of paraphrase knowledge for downstream semantic understanding tasks. Meanwhile, our proposed reconstruction method for the paraphrase corpus is independent of the specific language, which can provide more training sets for many languages and fields, generate high-quality paraphrase data, and further improve the performance of other tasks.
Key words paraphrase corpus construction; data augmentation; transfer learning; paraphrase recognition; natural language inference

自然语言处理的关键在于理解语言的歧义性和多样性，语言多样性指的是“异形同义”现象，即复述。复述是指同一语言中对同一意思的不同表达。由于人思想表达的复杂性导致实际的语言现象十分多样，机器处理自然语言任务时，面对同一种语义具有的多样化表达方式，难以准确理解语义并作出相同的反应，因此难以达到机器理解自然语言的目标。复述可广泛应用于复述生成[1]、机器翻译[2]和信息检索[3]等自然语言处理任务中，目前这些任务多采用神经网络模型框架，包含语言多样性的数据可以增强模型对语言现象的学习能力。因此构建复述语料，可以覆盖更多语言现象，从而提升模型鲁棒性。同时，如何利用复述数据以及验证复述对提升自然语言处理任务性能具有重要的意义。

另一方面，不同的语种具有不同的语言现象和特点，英语语法结构有较为严格的约束，日语的语法比英语更为灵活，但也具有格助词等直接的表层特征可以用于机器识别。而汉语相较于英语和日语有更加复杂的形态，也缺乏可以帮助机器识别的标志，因此加大了计算机理解汉语语义的难度。目前复述相关的研究及资源大部分集中与英语和日语，基于汉语的复述平行语料十分匮乏，阻碍了汉语复述研究的发展。我们提出利用多翻译引擎构建汉语复述平行语料的方法，将已有的英文复述语料迁移到汉语上，第一次在汉语上构建大规模高质量的复述平行语料，为汉语复述研究提供数据基础。

为验证所构建复述数据在不同自然语言处理任务上应用的有效性，我们分别将复述数据应用于复述识别任务和自然语言推理任务上。在基于注意力机制的句子匹配模型上进行实验和评测。根据迁移学习的思想，将预训练复述识别任务将学习到的复述知识迁移到自然语言推理任务中，并在公开的推理评测集上进行评测，验证复述数据对于语义理解任务的有效性。我们构建的复述语料，作为一种数据增强手段，可以弥补汉语其他任务训练样本不足的问题，增强自然语言处理任务模型的泛化能力。

本文其余部分组织如下：第二节介绍相关的研究和工作；第三节介绍构建大规模汉语复述平行语料的方法；第四节介绍复述有效性验证实验的设置；第五节介绍构建的汉语复述数据在复述识别上的应用；第六节介绍汉语复述语料应用于自然语言推理任务上的评测实验以及结果分析；第七节对本文研究进行总结。

1 相关工作

复述研究广泛应用于自然语言处理任务中，可以用于扩展信息检索和自动问答的问题[4]，也可以帮助扩展机器翻译的训练数据[2]，并且还可以应用于计算机辅助阅读[5]和自动摘要[6]等应用中，是自然语言处理领域的研究热点。在基于神经网络模型训练的技术主流下，复述数据的获取和构建对于复述研究十分重要。通过对复述相关研究的调研，我们发现国内外复述研究主要集中在英语和日语，汉语的复述生成研究相对滞后，可利用的汉语复述平行语料更是匮乏。与此同时，随着机器翻译技术的相对成熟，已有研究人员提出基于多种翻译引擎的多枢轴方法[7]，将源语言翻译成多个语言的译句，再通过反向翻译为源句得到多个候选复述，并利用不同的方法选取最佳的复述句对。这种方法能够获取更多样化的句子表达形式，增加复述的多样性可能。并且 Wieting 等人[8]提出反向翻译的方法，结合神经机器翻译的编码-解码框架，将源句经过二次翻译的译文作为输入句子的复述。同时与自然语言处理应用相结合，将构建的复述句对应用于句子语义表示的学习。因此利用机器翻译技术在一定程度上能够将丰富的英语复述平行资源迁移到汉语复述

语料，对推动汉语复述研究的发展有着重要意义。

对于其他自然语言处理任务中训练样本不足的问题，有研究人员通过同义词替换、随机插入删除或随机替换[9]等添加噪声的方法来进行文本数据增强，这些方法能够促进生成对抗网络模型[10]的发展，但同时会需要更多的数据资源，也会增加模型训练的难度。而我们的方法能够快速生成高质量的复述数据，并且复述知识可以增强其他任务的语义理解能力，应用于不同自然语言处理任务中能够提升性能，具有有效的数据增强能力。

2 汉语复述平行语料构建方法

2.1 基于多翻译引擎构建汉语复述平行语料

汉语复述研究的资源十分匮乏，导致神经网络模型的训练样本数量不足，难以利用复杂的深度学习模型来学习句子的语义特征。受到已有的数据增强方案启发，基于英语复述资源较为充足，神经机器翻译技术已经有十分成熟应用，以及主流的翻译引擎可以取得良好的翻译性能的现状，我们提出基于多翻译引擎的方法利用英语复述资源增强汉语复述数据。

我们采用由问句复述句对组成的 Quora 英语复述平行语料，由于不同翻译引擎在不同领域中表现出不同的翻译质量，我们首先在谷歌翻译、必应翻译、百度翻译、搜狗翻译和有道翻译这 5 个主流翻译引擎中，选择更适合 Quora 数据集的翻译引擎。为探究不同翻译引擎翻译各类句长的句子的效果，我们从 Quora 数据集中挑选 40 对句子用于翻译引擎筛选，包括句长为 5、10、15 和 20 词的复述句对各 10 对，分别经五个翻译引擎进行翻译。然后人工制定判断两个句子是否互为复述及其多样性的评分标准，由于人工标注存在标注者主观性强的问题，因此我们所制定的评分标注，尽可能通过量化的方法将句子的样式变化作出定义，以减小由于标注者不同而引起的统计结果偏差。具体评分标准如表 1 所示。

表 1　评测汉语译文句对的评分标准

Table 1 Scoring criteria on Chinese translation pairs

分数	评测标准
1	两个汉语译文翻译不准确，存在用词和语法错误超过 2 处。
2	两个汉语译文翻译部分不准确，存在用词和语法错误约 1-2 处。
3	两个汉语译文翻译基本正确，无语法错误，语法结构基本不变，词汇变化约 1-2 处。
4	两个汉语译文翻译准确且语法正确，语法结构基本不变而词汇变化超过 2 处；或语法结构存在简单变换。
5	两个汉语译文翻译准确，语法正确，语法结构存在较大变换；或词汇变化超过 3 处。

其中 1 分和 2 分判定为不可用的翻译结果，是非复述句对；3 分、4 分和 5 分表示是复述句对，并且表达方式更多样则人工评测得分越高，表明复述的质量越高。

采用表 1 评分标准对上述 40 对汉语译文进行人工评分，统计不同句长句对的译文评分结果在 3-5 分的个数。统计结果如图 1 所示。

统计结果表明，搜狗和有道翻译引擎的翻译性能在 Quora 数据集上优势最为明显。二者能够获得更多得分在 3-5 分的复述句对。此外，对于长度为 15 和 20 的长句，二者也生成了远超其他引擎的复述句对。

因此我们选取搜狗翻译和有道翻译来翻译 Quora 训练集中的每一对英文复述句，来构建汉语复述训练集。我们过滤掉了编辑距离小于 2 的翻译句对，即表现形式差异小的复述句对。由此分别得到了搜狗和有道的 13w 和 13.3w 的汉语复述句对，如将它们合并可以得到 26.3w 对的汉语复述训练集。采用同样的处理方式，我们对验证集和测试集分别翻译和过滤，得到了 1w 对汉语句对作为验证集，以及 1w 对测试集。

我们提出的基于翻译引擎的复述语料构建方法，在迁移数据方面具有通用性，我们的方法不只可以应用于 Quora 的英文问句数据集，同样也可以应用于陈述句或其他复述数据集。机器翻译引擎的应用十分成熟，并且英文的复述资源较为丰富，当机器翻译引擎具备英文至其他语言的翻译条件时，我们的方法可以为其他语言构建复述数据集提供可行的思路。

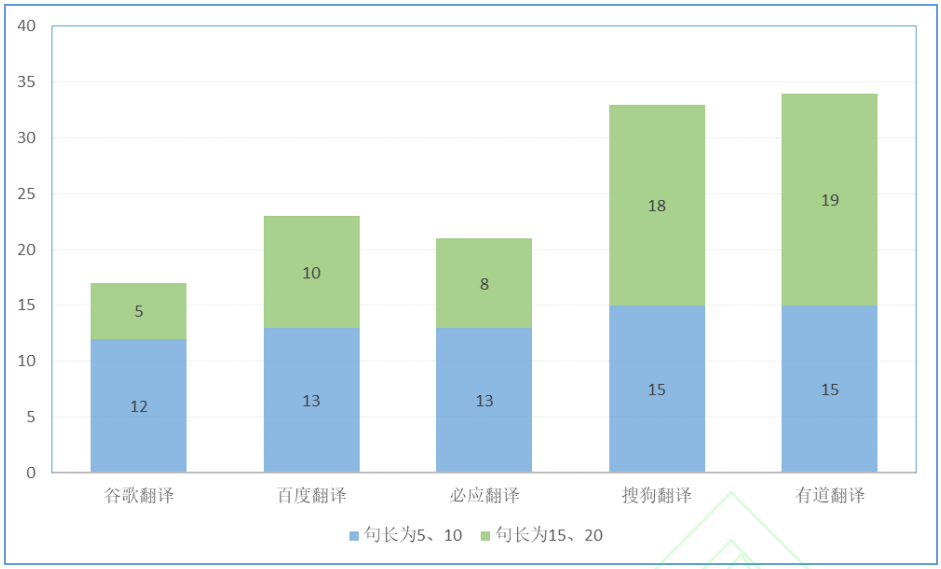

图 1　五个翻译引擎不同长度译文的评分统计结果

Fig. 1　Statistical results of scoring different-length Chinese translation pairs on the five translation engines

2.2 人工标注辅助方法构建复述评测集

由于机器翻译引擎方法得到的文本会有一定的损失，可能会存在汉语译文句对语义不一致的情况，因此为了构建一个高质量的汉语复述评测集，我们对译文进行了人工标注。数据集标注人数为 2 人，按照表 1 中定义的评分标准进行标注。具体来说，我们对获得的 1w 对汉语复述测试集进行注释，对每对符合 4 分或 5 分标准的翻译句对，筛选出作为复述评测数据。通过这种方式，我们过滤了翻译错误、语义不同、表达形式差异小的句对。最终，筛选后得到的高质量汉语复述句对，作为汉语复述评测集。构建的汉语复述评测集可应用于复述生成及相关任务的评测。

我们利用提出的方法所构建的汉语复述数据集及复述识别数据集①在开源网站中进行发布。

3 神经网络句子匹配模型

我们利用构建的复述数据，分别应用于复述识别与自然语言推理任务，进行复述数据的有效性验证。在复述识别任务上直接应用构建的复述数据作为训练集；在自然语言推理任务上，采用迁移学习[11]的思想，迁移在基于注意力机制的句子匹配深度网络模型上训练复述识别任务的模型参数。对自然语言推理任务模型进行微调后，进行自然语言推理任务的训练和评测，通过公开评测集的验证评估我们所构建的复述数据的质量。

3.1 基于注意力机制的句子匹配模型

复述识别任务和自然语言推理任务都可以看作是一种文本分类任务，输入两个句子，输出两个句子的相关性标签。二者均利用平行句对用于模型的训练，不同任务输出的标签类型不同。复述识别任务的输出标签包括两类：非复述 0 和复述 1；自然语言推理任务的输出标签包括三类：蕴含(entailment)、矛盾(contradiction)和中立（neural）。我们采用 Duan 等人[12]提出的基于注意力机制的句子匹配神经网络模型（AF-DMN），在两个任务上进行分类实验。

AF-DMN 句子匹配模型主要结构包括三层：编码层、匹配层和预测层。编码层将输入的两个句子分别经过一层 Bi-LSTM 网络，得到句子的上下文语义表示。匹配层基于注意力机制将两个句子表示进行联系，经过 cross attention 层、fusion 层、self-attention 层和 fusion 层后，得到新的句子表示。注意力机制[13]可以计算句子中每一个词和另一个句子中词汇的相关性，即注意力权重。匹配层中的 cross attention 表示跨句子注意力机制，是计算两个不同句子的注意力权重；同理 self-attention 表示自注意力机制，是计算一个句子

① https://github.com/Wys997/Chinese-Paraphrase-from-Quora

内单词之间的注意力权重。而 fusion 层是将句子的上一层表示和得到带有注意力权重的句子表示进行拼接融合的过程。最后预测层采用池化[14]的方法融合语义信息以得到定长的句子表示，将提取得到的语义信息输入全连接网络层进行标签预测。

3.2 模型细节设计

我们采用腾讯 AI Lab 开源①的大规模高质量中文词向量作为预训练的词向量，维度为 200 维，在模型训练过程中不更新词向量。模型参数设置如下：设置 Batch 大小为 16，在层之间采用了 Dropout 正则化技术，设置 drop 率大小为 0.35。我们采用 Adam[15]优化算法训练模型，设置初始学习率为 0.0004。在两个任务上我们都采用准确率作为评测指标来进行评测。

表 2　复述识别任务和自然语言推理任务数据统计信息

Table 2　Data statistical information for paraphrase recognition task and recognizing textual entailment task

数据集	训练集（对）	验证集（对）	测试集（对）	词汇表
复述识别（随机数据）	26w	1w	1w	2w
复述识别（翻译数据）	10w	1w	1w	2w
自然语言推理	9w	1w	1w	5k

4 复述识别任务上的验证实验

4.1 复述识别数据构建

我们基于Quora英文数据中标记为1的复述句对，已经获取了26.3w对汉语复述数据，其中搜狗引擎翻译结果和有道引擎翻译各有约13w对汉语复述句对。我们在此基础上构建汉语复述识别数据。其中训练集由两部分组成，以上述所述的13w搜狗翻译结果作为复述数据，即标记为1的正例样本。另一部分由非复述数据构成，标记为0作为负例样本。非复述数据我们采用两种方式来进行构建，一种是随机采样的方法，另一种为我们提出的基于翻译引擎的方法。最终，我们利用两种不同方法得到的训练集在复述识别任务上进行对比实验。

随机采样的方法是将有道翻译结果分为13个集合，为每个集合中的每一个原句在集合内随机匹配出目标句，组合的句对作为非复述数据。由于完全随机的匹配方式得到的非复述句对，其句子相似度几乎为0，训练集简单的分布会导致句子匹配模型对于字面相似却语义不同的非复述句对的学习能力不足。为提高模型的学习精度，我们在匹配目标句时通过计算句对的BLEU[16]值，选取BLEU值大于0的句对作为非复述句对，增加模型数据训练的难度，令模型学习更多的非复述句特征，共获得13w对非复述句对。其次是利用我们提出的方法基于搜狗翻译引擎翻译将Quora英文训练集中非复述数据句对翻译为汉语，得到5w对汉语非复述句对。训练集中保持正例和负例的数量相等，并且复述识别验证集和测试集的构建方式和训练集相同。最终构成的复述识别数据集的统计信息如表2所示。

4.2 实验过程及结果分析

我们利用不同的复述识别训练集分别来训练复述识别任务，模型参数均采用随机初始化的方法，所得实验结果与 AF-DMN 模型在 Quora 英文复述识别数据上的评测结果进行比较，英文数据实验结果来自模型复现的实验结果。实验结果如表 3 所示。

表 3　复述识别任务评测结果

Table 3 Experiment results on paraphrase recognition task

实验数据	Test Acc
Quora 英文数据	87.20
Quora 汉语随机数据	**50.00**
Quora 汉语翻译数据	**79.25**

① https://ai.tencent.com/ailab/nlp/embedding.html

从表 3 中的测试集准确率结果对比可以看出，我们构建的汉语随机数据用于复述识别任务的训练，所得结果距离英文数据评测结果具有较大的差距，并且测试集只达到 50 的准确率。这是由于我们构建的训练集的负例样本来自机器的计算和随机匹配，导致模型对于汉语数据中负例样本的学习难度较低，对正例样本的学习难度较高，因此模型难以泛化学习到复述的表示。利用翻译引擎方法获得的汉语数据用于复述识别任务的训练结果，虽然较英文数据实验结果有 7.95 的差距，但相比随机数据的结果有十分明显的提升。这说明我们的数据构建方法的有效性，利用我们提出的方法构建汉语复述平行语料能够有效迁移复述知识，并在复述识别任务上获得较高的性能。另一方面，在非复述数据的知识迁移表现上，我们的方法也比随机匹配方法更加有效。我们提出的基于翻译引擎的方法能够将复述和非复述中不同类型的数据分布信息迁移到汉语数据中，增强模型的学习表示和泛化能力，在复述识别任务上得到有效的验证。

我们调研汉语复述识别任务的实验结果，发现神经网络模型对于句子结构变化较大，但语义上确实为复述的句对会给出更高的非复述概率，难以正确识别为复述；而对于句子结构和词序相同或相似，却由于部分单词的不同导致语义不同的非复述句对，神经网络模型会在复述类别得到更高概率的结果。机器翻译的评测方法 BLEU 是基于文本语序的匹配来进行句子相似度计算，因此我们从上节所得的汉语复述测试集中抽取复述识别任务预测错误的句对计算 BLEU 值。我们总结发现复述识别任务难以识别的两类现象：高 BLEU 值的非复述句对，以及低 BLEU 值的复述句对，具体的示例如表 4 所示。

表 4　神经网络复述识别现象

Table 4　Phenomena on paraphrase recognition task

现象	示例	BLEU
高 BLEU 非复述	S：柬埔寨地震的主要影响是什么，这些影响与 **1963** 年**千岛**群岛地震相比如何？ T：柬埔寨地震的主要影响是什么，这些影响与 **1957** 年**安德里亚诺夫**群岛地震相比如何？	0.84
低 BLEU 复述	S：学习编码的**方法是什么**？ T：**如何**学习编程?	0.20

其中 S 和 T 分别代表输入数据对中的两个句子。从表中示例可以发现，示例 1 中的两个句子结构和语序相同，因为具体的时间和实体，“1963 年千岛群岛”和“1957 年安德里亚诺夫群岛”这两个部分不同导致两个句子的语义不一致，不能认为是复述，而模型预测为是复述。示例 2 中的两个句子结构差异较大，但都表达了“询问学习编程的方法”的语义，二者可以作为复述句，而模型预测为非复述。因此可以发现，神经网络复述识别模型针对上述两种类型的文本特征学习不足，难以正确识别，因此，构建更多具有多样性复述现象的数据十分重要，在复述识别任务上也能得到有效的应用。

5 在自然语言推理任务上的有效性验证实验

受 Wieting 等人[8]的工作启发，我们首先利用第五节中复述识别任务的模型，通过模型微调的方式，将学习到的模型参数迁移到自然语言推理任务中，在公开的推理数据集上评测，证明我们所构建的复述数据对改进自然语言推理性能的有效性。

5.1 实验数据

在自然语言推理任务上，我们使用的是CCL2018评测任务①中的汉语自然语言推理公开评测集，该评测集来自SNLI和MulitNLI两个数据集，经过机器翻译后再进行人工整理。CNLI自然语言推理评测集共有训练集9w对，验证集和测试集各1w对。数据统计结果如表2所示。

5.2 实验过程及结果分析

我们使用基于翻译引擎方法构建的复述识别数据训练 AF-DMN 模型所得的模型参数，除了最后一层参数不一样，其他层完全一致。我们在自然语言推理数据上微调模型，实验结果如表 5 所示。

① https://github.com/blcunlp/CNLI

表 5　自然语言推理任务评测结果

Table 5　Experiment results on natural language inference

模型	Dev Acc	Test Acc
可分解注意力模型[17]	69.35	—
基线模型	67.48	66.67
微调模型	**70.93（3.45↑）**	**69.31（2.64↑）**

其中基线模型是对所有参数均采用随机初始化的方法，利用 AF-DMN 模型训练及评测 CNLI 数据的实验结果。微调模型迁移 AF-DMN 模型在复述识别任务的模型参数，用于自然语言推理任务的数据集 CNLI 的训练和评测。其中可分解注意力模型为 CCL2018 评测任务比赛中公开的模型。

从表 5 可以看出，与基线模型相比，微调模型精度在验证集和测试集的准确率指标上分别提升了 3.45 和 2.64。利用我们所构建的复述数据训练模型，并经过模型微调后训练自然语言推理任务，相比于随机初始化参数进行模型学习有一定精度的提升。并且可以看出在验证集准确率上，微调模型超过了公开评测实验的结果，提升了 1.58 个点，说明我们所构建的数据使得实验能够在一定程度上得到超过公开的语料库基准的结果。上述实验结果说明我们构建的数据可以有效地用于复述识别模型的训练学习，并且复述知识经过迁移到自然语言推理任务中增强语义理解能力，提高模型学习精度，增加有监督学习方法的鲁棒性，由此证明我们所构建复述数据的质量以及多翻译引擎构建方法的作用。同时说明复述在自然语言处理任务中具有数据增强的作用，能够加强模型对语义的理解和学习能力，提升不同自然语言处理任务的性能。

表 6　小样本数据评测结果

Table 6　Experiment results on small sample data

数据量	模型	Test Acc
4.5w（50% CNLI）	基线模型	60.71
	微调模型	**62.94（2.23↑）**
3w （30% CNLI）	基线模型	50.69
	微调模型	**62.51（11.82↑）**

对于一些利用神经网络进行自然语言处理的任务来说，由于数据资源匮乏，数据量较少会导致模型学习性能较低。为验证我们所构建数据集用于小样本数据任务的质量，我们降低 CNLI 数据集的训练集和验证集的数据量，分别随机抽取原数据集中 50%和 30%的数据，构建小样本数据集。在基线模型和微调模型上进行对比实验，实验结果如表 6 所示。

我们在基线模型和微调模型上都分别利用两个小样本数据集进行训练和评测。从表 6 中可以看出，两个数据集在微调模型上训练和评测的结果，相较于基线模型结果分别有 2.23 和 11.82 的提升。实验结果说明，我们利用提出的利用翻译引擎方法所构建的复述数据，可以有效应用于复述识别任务，并且预训练模型参数后于自然语言推理任务上进行微调，对推理任务也有较大的提升。

另一方面，从实验结果中可以发现，当模型训练数据急速下降的时候，由原始的 9w 到 4.5w，最后降低到 3w 数据，基线模型和微调模型的性能都在降低。其中基线模型性能从 4.5w 数据到 3w 数据下降了约 10 个百分点，说明模型性能受训练数据数量的影响很大，数据量太小会大大降低模型性能。然而在微调模型中，在 3w 数据量时相较于基线模型有 11.82 个百分点的提升，约是 4.5w 数据量的提升值 2.23 的 5 倍。说明我们所构建数据预训练的模型微调后，能够有效提升低资源任务的性能。实验结果证明我们基于翻译引擎的方法所构建的复述数据的质量，能够在自然语言推理任务上得到验证，并且利用构建数据训练的微调模型，能够提高低资源任务的性能，针对自然语言处理下游任务的提升，我们构建的复述数据也具有十分重要的帮助和作用。

6 结语

本文基于自然语言理解中的瓶颈问题——语言表达的多样性，即复述现象，探索了复述语料对于自然

语言处理任务的重要性。本文提出一种基于多翻译引擎法构建复述平行语料的方法，在汉语复述的角度进行研究，将 Quora 英文复述数据迁移到汉语复述中，构建了大规模的汉语复述平行语料。其中训练集包括 26w 对复述句，验证集包括 1w 对复述句。并且根据获得的翻译结果，人工评测筛选得到高质量复述句的复述评测集。同时，基于构建的复述数据，分别应用于复述识别任务和自然语言推理任务，进行模型的训练和微调。实验结果表明我们构建的数据可以有效地用于复述识别模型的训练，在其上表现出与英文数据训练接近的性能。同时，复述知识经过迁移到自然语言推理任务中能够增强模型学习的能力，提高模型的鲁棒性，由此证明我们所构建复述数据的质量以及构建方法的有效性，也说明复述对改进自然语言处理其他任务的有效性。在低资源的领域和任务上，我们所构建的复述数据能够大大缓解语料不足导致模型性能低的问题，对自然语言处理其他任务的发展有积极的作用。

参考文献

[1] 赵世奇. 基于统计的复述获取与生成技术研究[D]. 哈尔滨工业大学.2009

[2] 姚振宇. 基于复述的机器翻译系统融合方法研究[D]. 哈尔滨工业大学. 2015

[3] Zukerman I, Raskutti B. Lexical query paraphrasing for document retrieval[C]// International Conference on Computational Linguistics. Association for Computational Linguistics, 2002:1-7.

[4] Mckeown K R. Paraphrasing using given and new information in a question-answer system[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1979:67-72.

[5] Carroll J, Minnen G, Pearce D, Canning Y, Devlin S, Tait J, Simplifying Text for Language-Impaired Readers, In Proceedings of EACL, 1999:269~270.

[6] Gambhir M, Gupta V, Recent automatic text summarization techniques: a survey, Artificial Intelligence Review, 2017, 47(1): 1~66.

[7] Kok S , Brockett C . Hitting the Right Paraphrases in Good Time[C]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA. Association for Computational Linguistics, 2010.

[8] Wieting J, Mallinson J, Gimpel K. Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 274-285.

[9] Wei J , Zou K . EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[J]. 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014. 2, 3, 4, 7

[11] Tan C , Sun F , Kong T , et al. A Survey on Deep Transfer Learning[J]. 2018.

[12] Duan C, Cui L, Chen X, et al. Attention-Fused Deep Matching Network for Natural Language Inference[C]// Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18. 2018.

[13] Bahdanau D, Cho K, Bengio Y, et al. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv: Computation and Language, 2014.

[14] Weng J , Ahuja N , Huang T S . Cresceptron: a self-organizing neural network which grows adaptively[C]// Neural Networks, 1992. IJCNN. International Joint Conference on. IEEE Xplore, 1992.

[15] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[16] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.

[17] Parikh A P , Tckstrm O , Das D , et al. A Decomposable Attention Model for Natural Language Inference[J]. 2016.