doi:10.3772/j.issn.2095-915x.2020.05.003

基于多翻译引擎的汉语复述平行语料构建方法

王雅松 刘明童 马彬彬 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

开放科学（资源服务）标识码（OSID）

摘要：复述指同一语言内相同意思的不同表达，复述生成指同一种语言内意思相同的不同表达之间的转换，是改进信息检索、机器翻译、自动问答等自然语言处理任务不可或缺的基础技术。目前，复述生成模型性能都依赖于大量平行的复述语料，而很多语言并没有可用的复述资源，使得复述生成任务的研究无法开展。针对复述语料十分匮乏的问题，我们以汉语为研究对象，提出基于多翻译引擎的复述平行语料构建方法，将英语复述平行语料迁移到汉语，构建大规模高质量汉语复述平行语料，同时构建有多个参考复述的汉语复述评测数据集，为汉语复述生成的研究提供一定的基础数据。基于构建的汉语复述语料，我们进一步对汉语复述现象进行总结和归纳，并进行复述生成研究。我们构建基于神经网络编码 - 解码框架的汉语复述生成模型，采用注意力机制、复制机制和覆盖机制解决汉语复述生成中的未登录词和重复生成问题。为了缓解复述语料不足导致的神经网络复述生成模型性能不高的问题，我们引入多任务学习框架，设计联合自编码任务的汉语复述生成模型，通过联合学习自编码任务来增强复述生成编码器语义表示学习能力，提高复述生成质量。我们利用联合自编码任务的复述生成模型进行汉语复述生成实验，在评测指标 ROUGE-1、ROUGE-2、BLEU、METEOR 上以及生成汉语复述实例分析上均取得了较好性能。实验结果表明所构建的汉语复述平行语料可以有效训练复述生成模型，生成高质量的汉语复述句。同时，联合自编码的汉语复述生成模型，可以进一步改进汉语复述生成的质量。

关键词：复述语料构建；汉语复述现象分类；复述生成；多任务学习；自编码任务

中图分类号：G35

基金项目：国家自然科学基金 (61876198, 61976015, 61370130, 61473294)，北京市自然科学基金 (4172047) 和科学技术部国际科技合作计划 (K11F100010)。

作者简介：王雅松（1996-），硕士，研究方向：自然语言处理、复述生成；刘明童（1993-），博士，研究方向：自然语言处理、神经机器翻译、复述；马彬彬（1993-），硕士，研究方向：自然语言处理；张玉洁（1961-），教授，研究方向：自然语言处理、机器翻译、文本大数据处理，Email：yjzhang@bjtu.edu.cn；徐金安（1970-），教授，研究方向：自然语言处理、机器翻译；陈钰枫（1981-），副教授，研究方向：自然语言处理、机器翻译。

doi:10.3772/j.issn.2095-915x.2020.05.003

The Construction of Chinese Paraphrase Parallel Corpus Based on Multiple Translation Engines

WANG Yasong LIU Mingtong MA Binbin ZHANG Yujie XU Jinan CHEN Yufeng

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Abstract: Paraphrases are sentences or phrases that express the same meaning using different wording. Paraphrase generation refers to generate different expressions with the same meaning in the same language, which is an indispensable basic technology to improve natural language processing tasks such as information retrieval, machine translation, automatic question answering, etc. At present, the performance of the paraphrase generation model relies on a large number of parallel paraphrase corpus. However, there are no available paraphrase resources in many languages, which makes the research on paraphrase generation task impossible. In view of the lack of paraphrase corpus, we propose a method of constructing large-scale and high-quality parallel corpus for Chinese paraphrase based on multiple translation engines to transfer parallel English paraphrase corpus to Chinese. At the same time, Chinese paraphrase evaluation datasets with multiple references are constructed to provide some basic data for the research on the generation of Chinese paraphrase. Based on the constructed Chinese paraphrase corpus, we further summarize and conclude the phenomenon of Chinese paraphrase and make research on paraphrase generation. We construct a Chinese paraphrase generation model based on neural network encoder and decoder framework, which adopt the attention mechanism, copy mechanism and coverage mechanism to solve the problems of unknown words and avoid word repetition in generation. In order to alleviate the problem of low performance of neural network paraphrase model caused by limited paraphrase corpus, we propose a Chinese paraphrase generation model with joint learning auto-encoding task. The model enhances the quality of paraphrase generation by improving the learning ability of encoder. We used the model to conduct the Chinese paraphrase generation experiment, and achieved good performance in both the quantitative evaluation on ROUGE-1, ROUGE-2, BLEU, and METEOR the analysis of generated Chinese paraphrase examples. Experimental results show that the constructed parallel Chinese paraphrase corpus can effectively train the paraphrase generation model and generate high-quality Chinese paraphrase sentences. Meanwhile, the quality of Chinese paraphrase generation can be further improved by Chinese paraphrase generation model with joint learning auto-encoding task.

Keywords: Paraphrase corpus construction; Chinese paraphrase phenomenon classification; paraphrase generation; multi-task learning; auto-encoding task

引言

复述生成作为自然语言处理的重要研究课题，旨在将给定的句子转换成多个语义相同但表达不同的句子，如给定原句 “who is the author of Hamlet?”，可以生成不同表达的句子（复述句）“who writes the novel of Hamlet?”、“who creates Hamlet?”。复述生成被广泛应用于机器翻译系统[1]、自动问答系统[2]、信息抽取[3]、信息检索[4]、自然语言生成[5]和自动文摘[6]等自然语言处理任务中，以提高系统对语言现象处理的覆盖面，增强鲁棒性。

目前基于神经网络的复述生成研究大部分集中在英语和日语上，汉语方面的研究与国际先进水平相差交大。汉语复述研究滞后的原因有两方面，一方面是汉语复述平行语料十分匮乏，严重阻碍了汉语复述研究。因此，探索高效构建汉语复述平行语料的方法成为首要任务。

另一方面是汉语有着比英语和日语更为复杂的语言现象，比如英语有较严格的语序约束，日语有丰富的格助词等表层信息，而汉语在语序上自由，同时缺少形态标志，增加了分析和生成的难度。英语和日语上的复述研究都在早期完成了各自语言中复述现象的语言学分类，为其复述生成研究奠定基础。但是，在汉语方面依旧是空白，因此，开展并完成在汉语中复述现象的语言学分类研究成为当务之急。考虑到英语具有较为丰富的复述平行语料，并且机器翻译技术比较成熟，我们提出基于多翻译引擎的汉语复述平行语料构建方法，并开展汉语复述现象分类研究和汉语复述生成研究。

随着编码 - 解码神经网络在端到端文本生成任务中的成功应用，研究人员开始将编码 - 解码框架应用于复述生成，利用编码器将原句的上下文编码为语义向量，之后利用解码器将该语义向量解码为对应的复述句，使得编码器和解码器共享同一个语义空间。基于编码 - 解码框架的复述生成主要面临两方面问题：一方面，生成的复述句中存在生成不准确以及生成词汇重复的问题；另一方面，规模有限的复述平行语料导致编码器的语义表示学习效果尚且不足 [7-9]。如何有效利用有限的复述平行语料，提升复述生成质量成为研究的重点。针对已构建的汉语复述语料，我们采用联合自编码任务的多机制融合复述生成模型，引入多任务学习框架，联合学习复述生成任务和自编码任务，通过两个任务共享同一个编码器，共同提高编码器的语义表示学习能力，从而提高汉语复述生成的质量。

本文其余部分组织如下：第一节介绍相关研究；第二节介绍大规模汉语复述平行语料的构建方法；第三节介绍汉语复述现象分类研究；第四节介绍联合自编码任务的复述生成模型的实现；第五节介绍在已构建语料上的评测实验和结果分析；第六节对本文研究进行总结。

1 相关研究

现有的复述生成研究都依赖于大规模的复述平行语料，而已有的高质量复述语料十分匮乏，复述数据的获取和构建对于复述生成及其相关研究十分重要。已有的复述获取采用基于枢轴的方法 [10]，利用双语平行语料，双语平行语料是机器翻译的训练语料，语义相同的线索隐藏在另一个语言中，如果两个短语对齐到另一语言的同一短语上，则二者互为复述。但这种方法得到的复述表达方式的变化较小，且需要缓解单词对齐带来的噪声问题，所以方法的效果十分有限。也有研究人员提出基于机器翻译引擎的复述句获取方法，将二次翻译的译文作为输入句子的复述 [11]。Mallinson 等 [12] 将神经机器翻译编码 - 解码框架结合双语枢轴方法，对双语句对进行逆向翻译得到复述句对。通过对复述生成研究的调研，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，阻碍了汉语复述生成研究的发展。因此研究如何构建汉语复述平行语料的高效方法对推动汉语复述研究有着重要意义。考虑到英语具有丰富的复述平行语料以及机器翻译技术相对成熟，我们提出基于多翻译引擎的汉语复述语料构建方法，迁移英语复述数据，同时筛

doi:10.3772/j.issn.2095-915x.2020.05.003

选得到高质量的译文，构建高质量的汉语复述平行语料，开展汉语复述的研究。其次，有研究人员调研了英语的复述现象[13]，比如主动态和被动态的替换，动词的过去时和现在时的替换，而汉语的复述现象分类和总结的工作还未有相关研究。我们在构建的汉语复述平行语料上，开展汉语复述现象研究和汉语复述生成研究。

在基于编码 - 解码神经网络框架[14]的复述生成方面，研究人员提出了许多改进方案。Prakash 等[15]通过在编码 - 解码框架的 LSTM 网络层间加入残差网络，实验结果超过了基于注意力机制的编码 - 解码模型的性能。Gupta 等[8]提出变分自编码器（VAE）与 LSTM 相结合的框架，能够生成多个对应的复述句。Zichao Li 等[7]将强化学习应用于编码 - 解码框架中，进一步改进了复述生成的质量。Brad 等研究人员[9]利用迁移学习，将训练的文本蕴含生成模型参数迁移到复述生成任务中，缓解复述平行语料不足的问题。这些神经网络复述生成的研究主要集中在英语上，并未在汉语上展开。如何利用神经网络模型，研究汉语复述生成是亟需解决的问题。基于构建的汉语复述平行语料数据集，本文进一步研究基于神经网络的汉语复述生成方法。

2 汉语复述平行语料构建

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，严重阻碍了汉语复述生成研究的发展。因此研究如何构建汉语复述平行语料的高效方法对推动汉语复述研究有着重要意义。考虑到英语具有丰富的复述平行语料以及机器翻译技术相对成熟，我们提出基于多翻译引擎的汉语复述语料构建方法，将 Quora 英语复述平行语料迁移到汉语。并在构建的汉语复述平行语料上，开展汉语复述现象研究和汉语复述生成研究。

2.1 基于多翻译引擎的汉语复述平行语料构建方法

基于神经网络的机器翻译技术不断成熟，包括谷歌翻译、必应翻译、百度翻译、搜狗翻译和有道翻译在内的主流翻译引擎取得良好的翻译性能。但不同的翻译引擎对不同领域的数据翻译质量参差不齐。Quora 英语平行复述语料由问句复述句对组成，为了选择更适合 Quora 数据集的翻译引擎，我们从 Quora 数据集中分别挑选句子长度为 5、10、15 和 20 词的复述句对各 10 对，利用上述五个翻译引擎分别进行翻译。然后综合考虑语义忠实度、语法正确度、表达流畅度和是否互为复述四个方面制定人工评分标准，对五个翻译引擎所得到的汉语译文句对进行评分。具体评分标准如表 1 所示。

其中 1 分和 2 分表示汉语译文翻译不准确，翻译结果不可用，且难以构成复述句对；3 分、4 分和 5 分表示在翻译正确的基础之上，主要从是否为复述句对以及表达方式的多样性两方面对汉语译文句对进行评测，得分越高，句子的变化形式更加多样，复述的质量越高。

本文采用以上评分标准对上述 40 对不同长度的汉语译文进行人工评分，由五个翻译引擎获得的汉语译文的人工评分结果统计如图 1 所示。其中，纵坐标表示不同句长的句对译文评分结果在 3~5 分的个数统计值。

表 1　评测汉语译文句对的评分标准

分数	评测标准
1	两个汉语译文翻译不准确，存在较多语法错误。
2	两个汉语译文翻译部分不准确，语法错误较少。
3	两个汉语译文翻译基本正确，无语法错误，表达形式变化单一。
4	两个汉语译文翻译准确，语法正确，表达形式变化较为多样。
5	两个汉语译文翻译准确，语法正确，表达形式变化丰富多样。

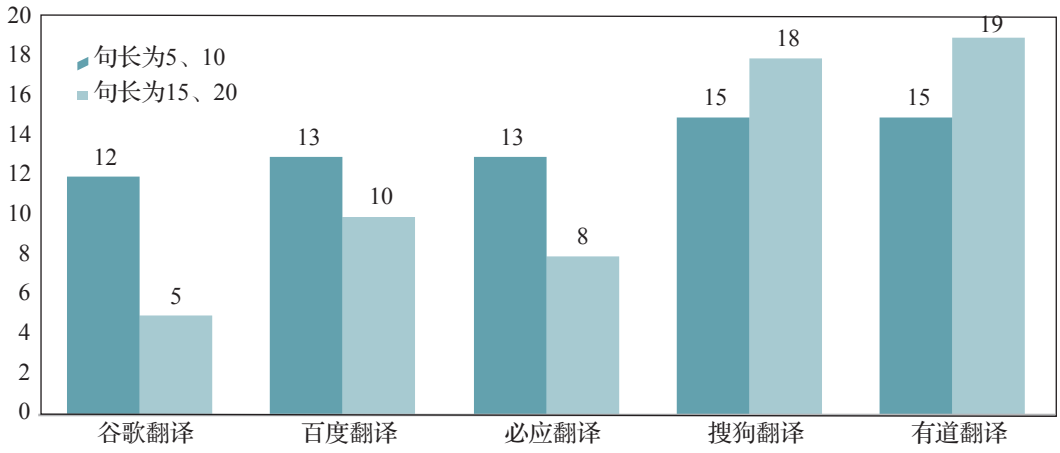

图 1　五个翻译引擎不同长度译文的评分统计结果

对句长为 5 和 10 的短句子来说，五个翻译引擎评分为 3~5 分的统计结果性能相当，搜狗和有道翻译有略微优势。对于句长为 15 和 20 的长句子，搜狗和有道评分为 3~5 分的句对分别有 18 个和 19 个，翻译性能优势明显。综合不同句长的评分统计结果，五个翻译引擎的综合排名为：有道、搜狗、百度、必应、谷歌，有道和搜狗翻译引擎可以在 Quora 数据集上取得更好地翻译性能。

然后，我们为汉语复述生成模型的训练集构建汉语复述平行语料。我们采用有道翻译和搜狗翻译两个翻译引擎对 Quora 数据集中每个英语复述句对进行自动翻译，分别得到训练集汉语复述句对 139k 和 139k。由于英语复述句对翻译后可能会得到一对相同的汉语译文，我们利用编辑距离对有道翻译和搜狗翻译两个翻译引擎的译文进行预处理，过滤编辑距离小于 2 的句对，分别得到 130k 和 133k 的汉语复述句对，最终构建得到汉语复述训练集 263k 对。同理，对 Quora 开发集进行翻译，对译文句对预处理之后得到 10k 对汉语复述开发集。

2.2　人工标注构建复述评测集

为了构建标准汉语复述评测集，我们对翻译 Quora 英文测试集得到的汉语句对进行人工标注。具体来说，我们利用有道和搜狗翻译引擎翻译 Quora 测试集的 5k 句对，整合两个翻译引擎的译文，共得到汉语句对 10k 对；然后根据人工定义的标准，进行人工标注。标注的标准主要从语义一致性和表达方式多样性两个

doi:10.3772/j.issn.2095-915x.2020.05.003

方面判断汉语句对是否互为复述。我们筛除翻译结果不正确、汉语译文在语义方面不构成复述和译文结果单一的句对。最终，我们共得到1226对高质量的汉语复述句对，作为汉语复述评测集。

2.3 汉语复述多参考评测集构建

高质量的复述句不仅语义一致而且表达方式多样，而已有复述生成评测方法通常利用单个参考复述进行评测，导致生成的复述往往由于表达方式多样而与参考复述相差较大，难以获得高的评测结果。我们考虑增加参考复述，尽量覆盖自动生成可能有的复述现象。因此我们在2.2节构建的评测集基础之上，构建带有多个参考复述的评测集。

考虑到不同翻译引擎对同一个句子的翻译结果在句法结构、表达顺序和翻译风格等各方面存在差异，不同引擎的翻译结果之间可互为复述，因此我们采用跨翻译引擎的方式构建带有多个参考复述的评测集。给定一对英语复述句对，本文采用搜狗翻译和有道翻译两个翻译引擎构建四个汉语复述句子，然后对每一个待评测的复述句构建三个参考复述。具体地，将2.2节由搜狗翻译引擎得到的汉语复述句对中所有的原句子作为多参考复述集的原句，而复述句子以及有道翻译引擎得到的汉语复述句对中的原句子和复述句子分别作为三个不同的参考复述集合。通过对2.2节所得高质量汉语复述对进行多样性筛选，筛除两个翻译引擎的翻译结果相同的句对，共得到带有三个参考评测集250对。

综合上述构建的汉语复述平行语料，具体数据统计如表2所示，其中单个参考评测集指的是2.2节构建的评测集。

表2 汉语复述平行语料

训练集	开发集	单个参考评测集	带有多个参考评测集
260k对	10k对	1.2k对	250对

3 汉语复述现象的分类

由于汉语和英语有着众多不同的语言特征，常见的英语复述现象并不能完全符合汉语复述现象，如英语可通过时态转换呈现不同的表达方式，而汉语语言本身并没有明显的时态特征。根据汉语本身的语言特征，我们在第2节构建的大规模汉语问句复述语料上，参考前人对英语复述现象的调研方法[13,16,17]，分析大量的汉语复述实例，首次对汉语复述现象进行分析总结。

在对汉语的复述现象分类时，我们所采用的分类粒度和Bhagat等[13]的略有不同，如Bhagat等将名词与动词的互换、动词与形容词的互换、动词与副词的互换、动词与语义角色名词的互换分别列为4种不同的复述现象，而我们将此类词性之间互换的现象统归为“词性变换”这一类。

根据汉语的语言特征，我们共总结出13种常见的汉语复述现象，总结如下：

（1）同义词（短语）替换：句对之间的变化主要体现为同义词或同义短语的替换。

（2）语序变换：原句中的某些成分的位置在复述句中发生变化。如由于表达习惯或写作风格的不同，时间状语、地点状语的位置会出

现在句首或句中或句尾。

（3）词典注释替换：原句中的词在复述句中被替换成对应的词典注释。词典注释替换的复述现象通常在专业领域文本中较为普遍，将晦涩的专业术语复述为通俗易懂的注释便于读者理解。

（4）词性变换：原句中的词或短语在复述句中被替换为对应的其他词性，如将原句的名词转换为对应的动词等。

（5）句子结构变换：原句和复述句整体的表述结构有较大的变动。该种复述现象变换较为复杂，并不是简单的个别成分替换或变化，有时包含多种复述现象。

（6）句子拆分与合并：句子拆分指原句被复述句拆分成多个简单句；句子合并与句子拆分过程相反。

（7）把字句变换：原句中的主动语态和被动语态表达在复述句中被替换为“把字句”结构。

（8）被字句变换：原句中的主动语态在复述句中被替换为“被字句”结构。

（9）对立关系互换：原句中的“a 对 b 具有某种关系 P”在复述句中被替换为“b 对 a 具有某种关系 Q”，其中 P 和 Q 互为对立关系，如原句中“a 比 b 小”在复述句中被替换为“b 比 a 大”等。

（10）否定语义 / 肯定表达互换：原句中含有否定意思的短语或片段在复述句中以肯定的方式进行表达，如原句中“没有债务”在复述句中被替换为“债务为零”等。

（11）重复和省略互换：句对之间的变化主要体现在某些片段的省略和重复。

（12）外部知识引入：原句中的某些词或短语在复述句中被替换为某些基于外部知识的词或短语。通过引入外部知识补充信息，保证语义一致。

（13）直述和间述变换：原句中的直接叙述方式在复述句中被替换为间接叙述的方式，反之亦然。

为了更好地理解汉语复述现象，我们给出每种复述现象的对应复述实例，如表 3 所示。

通过与已有的英语复述现象分类进行对比，我们发现“时态变换”和“细微变换”是英语特有的两种复述现象，而“把字句变换”、“被字句变换”和“否定语义 / 肯定表达互换”3 种复述现象为汉语特有。虽然本文汉语复述的分类粒度和英语复述研究不尽相同，但是除去英语和汉语各自特有的复述现象，本文总结的 13 种常见汉语复述现象基本涵盖了前人研究总结的 25 种英语复述现象类型，一定程度上佐证了本文汉语复述现象总结的完整性和科学性。我们在大规模汉语复述平行语料上所进行的汉语复述现象的总结将对推动汉语复述研究开展具有一定的参考价值。

4 联合自编码任务的多机制融合复述生成模型

本文采用联合自编码任务的多机制融合复述生成模型，采用融合注意力机制、复制机制和覆盖机制的编码 - 解码框架进行复述生成。将一个句子的复制看作该句的特殊复述句，作为自编码任务加入复述生成任务中。由此，通过两个任务共享同一个编码器，联合学习并提高编码器的语义表示学习能力。

doi:10.3772/j.issn.2095-915x.2020.05.003

表 3　汉语复述现象分类及实例

序号	类型	汉语复述实例
1	同义词替换	怎么在quora上谋生呢? / 如何通过quora赚钱?
2	语序变换	什么武术最适合自卫? / 自卫最好的武术是什么? 谁会成为印度的下一任总统? / 谁能成为下一任印度总统?
3	词典注释替换	比尔·盖茨捐了这么多钱，他怎么还能跻身世界前三? 比尔·盖茨捐出这么多钱，他怎么还是世界上最富有的三个人之一?
4	词性变换	我怎样才能说一口流利的英语? / 我如何流利地说英语? 禁止发行1000和500卢比纸币将如何影响印度经济? / 1000卢比和500卢比的禁令将如何影响印度经济?
5	句子结构变换	变漂亮有什么秘诀吗?/怎么能漂亮呢?
6	句子拆分与合并	如果忘记密码，我怎么能解锁我的iPhone? / 如何在没有密码的情况下解锁iphone?
7	把字句变换	为什么一些北印度人把v读成b ?/为什么v在印度北部发音为b ?
8	被字句变换	为什么发明比基尼? / 比基尼为什么要被发明?
9	对立关系互换	如果丈夫比妻子小5岁，将来会面临什么问题?/如果我妻子比我大5岁，将来会有什么问题?
10	否定语义/肯定表达互换	有没有没有国家债务的国家? / 有没有国家债务为零的国家? 反堕胎法是否禁止非贫困妇女堕胎? /反堕胎法是否禁止穷人以外的人堕胎?
11	重复和省略互换	学机械工程专业好，还是学计算机科学专业好? /学机械工程和计算机科学哪个专业好?
12	外部知识引入	苹果什么时候发布新款macbook?/苹果科技公司什么时候发布新款macbook?
13	直述和间述变换	上帝说：“我将永远与你们同在”，这句话可信吗? /上帝说他将永远与我们同在，可信吗?

4.1　多机制融合的复述生成模型

为了解决已有编码 - 解码复述生成方法中存在的问题，我们采用在基于注意力机制的编码 - 解码框架中融合复制机制和覆盖机制的方法，其中复制机制解决未登录词无法生成问题，覆盖机制解决词汇重复生成问题。

4.1.1　基于注意力机制的复述生成模型

基于注意力机制的复述生成模型采用长短时记忆循环神经网络 LSTM 作为编码器和解码器。

其中，编码器依次读取原句 X=[x_1, …, x_n] 中的每个词并进行语义编码，生成原句语义表示，解码器根据原句语义表示逐词生成目标复述句。在解码过程中，注意力机制 [18] 根据当前时间步 t 时刻解码器的状态 s_t 和编码器状态 H=[h_1, …, h_n]，动态计算时间步 t 时对于每个隐藏层状态的注意力权重，其中注意力权重越大的词，在当前时刻被给予越多的关注，注意力机制的计算如公式 1 和 2。

$$e_i^t = V^T \tanh(W_h h_i + W_s s_t + b_{attn}) \quad (1)$$

$$a^t = softmax(e^t) \quad (2)$$

其中e_i^t表示当前时间步 t 时的解码器状态 s_t 与编码器状态 h_i 之间的对齐得分，W_h、W_s、b_{attn} 表示可学习的参数，a_i^t表示归一化后对原句中每个词 x_i 的注意力权重。

4.1.2　多机制融合模型的目标函数

融合复制机制 [19] 和覆盖机制 [20] 的复述生

成基线模型采用交叉熵损失函数作为训练目标函数，计算如下公式 3 所示。

$$Lose_{copy+coverage}(\theta)=\frac{1}{T}\sum_{t=1}^{T}\left(-\log P(w_t)+\lambda\sum_i \min(a_i^t,c_i^t)\right) \quad (3)$$

$$P(w)=P_{gen}P_{vocab}(w)+(1-P_{gen})P_{copy}(w) \quad (4)$$

公式 3 括号中第一项表示生成词语的损失函数，词的预测概率 $P(w_t)$ 通过公式 4 计算获得。其中，P_{gen} 表示当前词选择生成模式预测的概率，其计算如公式 5；$1-P_{gen}$ 表示当前词选择复制模式预测的概率。$P_{vocab}(w)$ 表示集内词 w 的生成概率，其计算如公式 6；$P_{copy}(w)$ 表示原句中词 w 的复制概率，其计算如公式 7。当词 w 为未登录词或低频词时，$P_{vocab}(w)=0$；当词 w 在原句中不存在时，$P_{copy}(w)=0$。

$$P_{gen}=\sigma(w_h^T c_t+w_s^T s_t+w_{y_{t-1}}^T y_{t-1}+b_{gen}) \quad (5)$$

其中 w_h，w_s，$w_{y_{t-1}}$ 和 b_{gen} 为可学习的参数，σ 表示 sigmoid 激活函数。

$$P_{vocab}=softmax(U(tanh(V[y_{t-1},s_t,c_t]+b_v))+b_u) \quad (6)$$

其中 U、V、b_v、b_u 表示可学习参数。

$$P_{copy}(w)=\sum_{i:w_i=w}a_i^t \quad (7)$$

其中 a_i^t 为时间步 t 时对每个词 x_i 的注意力权重。

公式 3 括号中的第二项表示覆盖机制的损失函数，其中 a_i^t 和 c_i^t 表示在 t 时刻对词 x_i 的注意力决策，其中 a_i^t 由公式 2 得到，c_i^t 表明记录历史决策注意力信息的覆盖向量，其计算由公式 8 得到。其意义可解释如下，若历史决策中对 x_i 单词的注意力权重比较大，则当前时刻 t 对该单词的注意力需适当减小；反之，若当前时刻 t 对该单词的注意力权重较大，则历史决策中对单词 x_i 的注意力权重需适当减小。通过惩罚重复关注同一词的注意力决策，以规避重复词的生成。而超参数 λ 用来对覆盖损失进行加权平衡。

$$c^t=\sum_{t'=0}^{t'=t-1}a^{t'} \quad (8)$$

其中 $a^{t'}$ 表示时间步 t' 时刻原句中各个词语的注意力对齐权重，其计算公式与公式 2 一样。但是，e_i^t 的计算进行了扩展，将覆盖向量 c^t 引入到注意力计算公式 1 中，如公式 9 所示，使得注意力决策受到历史决策信息的约束。

$$e_i^t=V^T tanh(W_h h_i+W_s s_t+W_c c_i^t+b_{attn}) \quad (9)$$

其中 W_c、V^T、W_h、W_s、b_{attn} 表示可学习的参数。

4.2 联合自编码任务的复述生成模型

4.2.1 复述生成任务和自编码任务的联合学习

考虑到两个相同的句子是特殊的复述句对，可以作为自编码任务对其编码和解码，在多机制融合模型中引入多任务学习框架，联合学习自编码任务和复述生成任务，整体框架如图 2 所示。

模型由一个共享的编码器和两个解码器构成，其中编码器 - 解码器 1 用于新引入的自编码任务，生成原句本身；编码器 - 解码器 2 用于一般的复述生成任务。如图 2 中的示例，给定原句“曹雪芹写了小说《红楼梦》”，编码器编码学习原句的语义表示，经过解码器 1 生成原句本身“曹雪芹写了小说《红楼梦》”。并且经过解码器 2 生成对应的复述句“曹雪芹是小说《红楼梦》的作者”。联合学习的目的在于两个任务共同学习同一个编码器，以相互增强编码器的语义表示学习能力。

doi:10.3772/j.issn.2095-915x.2020.05.003

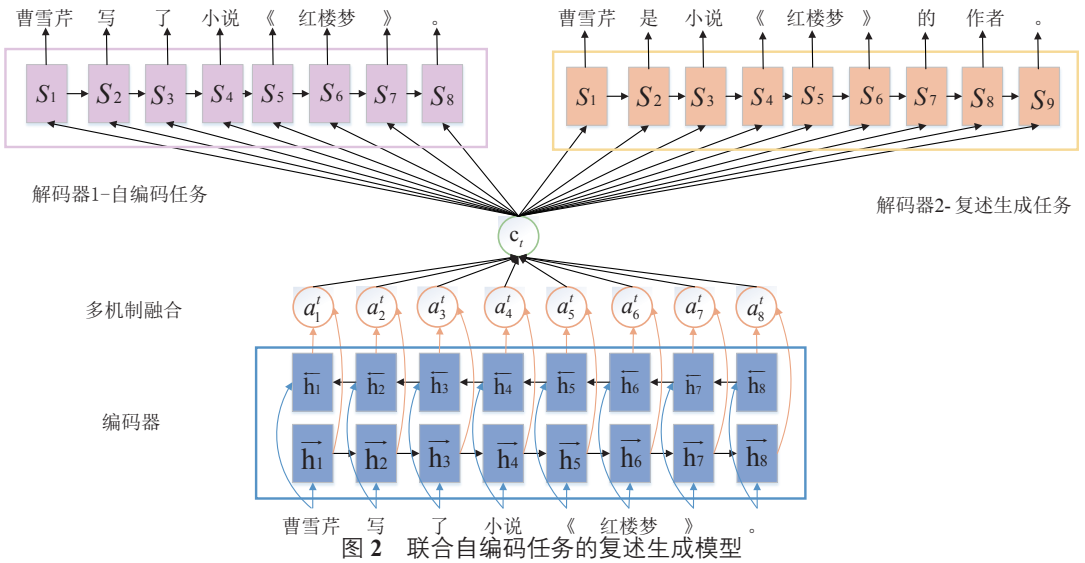

图 2 联合自编码任务的复述生成模型

4.2.2 联合学习的目标函数

本文融合自编码任务的复述生成模型采用交叉熵损失函数作为训练目标函数，模型的目标函数定义如下。

$$L(\theta)=\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{T_y}-logp(y_{i,n}\mid y_{1:i-1,n},X_n,\theta)+\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{T_y}-logp(x_{i,n}\mid x_{1:i-1,n},X_n,\theta) \quad (10)$$

其中第一项表示复述生成任务的损失，第二项表示自编码任务的损失。$\theta=(\theta_{enc},\theta_{dec_p},\theta_{dec_{auto}})$ 表示模型所有参数集合，θ_{enc} 表示编码器的参数集合，θ_{dec_p} 表示复述生成任务的解码器参数集合，$\theta_{dec_{auto}}$ 表示自编码任务的解码器参数集合。N 表示样本数，T_y 表示样本 n 的复述句 Y_n 中含有 T_y 个词，$Y_{i,n}$ 表示其中第 i 个词，T_x 表示样本 n 的原句 X_n 中含有 T_x 个词，$x_{i,n}$ 表示其中第 i 个词。

5 评测实验与结果分析

5.1 实验数据及评测指标

我们利用第2节构建的汉语复述平行语料，分别采用 4.1 节提出的多机制融合的复述生成模型和 4.2 节提出的联合自编码任务复述生成模型实现汉语的复述生成。

实验数据为构建的汉语复述语料，数据统计结果如表 2 所示。在数据预处理时，我们采用开源的 jieba 中文分词工具①对汉语复述语料进行中文分词。评测时，我们在 2.2 节构建的单参考复述评测集和 2.3 节构建的多参考复述评测集上，分别利用评测指标 ROUGE-1[21]、ROUGE-2、BLEU[22] 和 METEOR[23] 以及多个参考评测指标 ROUGE-1_{muli}、ROUGE-2_{multi}、$METEOR_{multi}$ 进行评测。其中多参考评测指标计算公式如下：

$$ROUGE-N_{multi}=argmax_i ROUGE-N(r_i,s) \quad (11)$$

$$METEOR_{multi}=argmax_i METEOR(r_i,s) \quad (12)$$

其中 s 为复述句原句，和所有参考句 r_i 的最高得分作为最终的评测结果。

5.2 评测实验结果及分析

5.2.1 基于多机制融合模型的汉语复述评测结果

本实验中，我们将基于注意力机制的复述生成模型作为基线模型，分别评测添加覆盖机制、

复制机制和多机制融合方法在已构建的汉语复述语料上的复述生成质量。评测结果如表 4 所示。

表 4　多机制融合模型的汉语复述生成评测结果

模型	ROUGE-1	ROUGE-2	BLEU	METEOR
基线模型	56.64	27.01	36.59	27.38
基线+复制机制	62.13(5.5↑)	32.38(5.3↑)	41.60(5.1↑)	31.22(3.9↑)
基线+覆盖机制	56.75(0.1↑)	27.98(0.9↑)	37.85(1.3↑)	27.96(0.6↑)
基线+复制机制+覆盖机制	61.97(5.3↑)	32.42(5.4↑)	42.04(5.5↑)	31.39(4.0↑)

通过分析表 4，我们可以看到，在 ROUGE-1、ROUGE-2、BELU、METEOR 指标上，与基线模型相比，融合覆盖机制、复制机制和多机制融合的神经网络模型对复述生成质量的提升有不同程度的贡献，充分验证了各个机制以及多机制融合的复述生成模型在汉语复述生成质量提升的有效性。同时，在人工制定的评测数据集上的实验结果也表明，我们通过基于多翻译引擎的跨语言复述知识迁移，可以构建高质量的汉语复述平行语料，从而有效训练汉语复述生成模型。

此外，我们在人工构建的汉语复述多参考评测集上进行了多机制融合模型的评测实验，实验结果如表 5 所示。

表 5　多机制融合模型的汉语复述多参考评测集结果

模型	ROUGE-1_{multi}	ROUGE-2_{multi}	BLEU	METEOR$_{multi}$
基线模型	46.75	21.59	52.90	32.29
基线+复制机制	51.77(5.0↑)	25.69(4.1↑)	58.58(5.6↑)	35.88(3.6↑)
基线+覆盖机制	47.11(0.4↑)	21.70(0.2↑)	54.22(1.3↑)	32.84(0.6↑)
基线+复制机制+覆盖机制	52.89(6.1↑)	26.44(4.9↑)	61.64(8.7↑)	36.58(4.3↑)

分析表 5，我们可以看到相较于基线模型，在 ROUGE-1_{muli}、ROUGE-2_{multi}、BLEU、METEOR$_{multi}$ 指标上，融合覆盖机制、复制机制和多机制融合对复述生成质量都有所提升，其中复制机制和多机制融合方法提升效果显著。由于基于多个参考的评测能覆盖更多可能生成的复述现象，评测结果更接近人工评测。因此，本实验结果更全面地验证了我们构建的汉语复述平行语料，能够有效训练汉语复述生成模型，并且训练得到的模型在人工构建的多参考评测集上的评测结果有效。

5.2.3　联合自编码任务的汉语复述评测结果

本实验采用正序解码和逆序解码两种不同的自编码任务解码方式进行对比实验，在单参考评测集上的评测结果如表 6 所示。其中 4.1 节的多机制融合的复述生成模型作为本实验的基线模型；Auto-Forward 为联合自编码的正序解码模型；Auto-Reverse 为联合自编码的逆序解码模型。

① https://github.com/fxsjy/jieba

doi:10.3772/j.issn.2095-915x.2020.05.003

表 6　联合自编码任务模型的汉语复述生成评测结果

模型	ROUGE-1	ROUGE-2	BLEU	METEOR
多机制融合模型	61.97	32.42	42.04	31.39
Auto-Forward	62.10(0.2↑)	33.45(1.0↑)	42.54(0.5↑)	31.50(0.2↑)
Auto-Reverse	62.88(0.9↑)	33.84(1.4↑)	43.07(1.0↑)	32.21(0.9↑)

相较于基线模型，联合自编码的正序解码模型在各个评测指标上都有所提升，并且联合自编码的逆序解码模型在各个指标上性能均超过正序解码结果，验证了联合自编码任务复述生成模型和逆序解码方式的有效性，说明自编码模型能够增强编码器的语义表示学习能力，在一定程度上能够补足复述语料匮乏导致的语义学习不足的问题。

同样地，我们利用带有多个参考的评测集进行评测，评测结果如表 7 所示。

表 7　联合自编码任务模型在多参考评测集上的评测结果

模型	ROUGE-1$_{multi}$	ROUGE-2$_{multi}$	BLEU	METEOR$_{multi}$
多机制融合模型	52.89	26.44	61.64	36.58
Auto-Forward	53.20(0.31↑)	26.87(0.43↑)	61.80(0.24↑)	37.06(0.48↑)
Auto-Reverse	53.59(0.70↑)	27.03(0.59↑)	62.23(0.59↑)	37.18(0.60↑)

分析表 7，我们同样可以看到相较于基线模型，在 ROUGE-1$_{muli}$、ROUGE-2$_{multi}$、BLEU、METEOR$_{multi}$ 上，联合自编码的正序解码模型都有较大提升，联合自编码的逆序解码模型相较于正序解码模型有更大的提升。多参考的评测结果充分验证了联合自编码任务的复述生成模型和逆序解码方式在提升汉语复述生成质量方面的有效性，以及我们构建的多参考汉语复述评测集对于评测复述生成模型的有效性。

为了更直观的分析自编码任务采用正序解码和逆序解码两种不同方式对生成复述质量的影响，我们在表 8 给出联合自编码任务的正序解码模型和联合自编码任务的逆序解码模型生成的汉语复述实例。

分析表 8 中例子 1，首先，我们可以看到，无论是联合正序解码还是逆序解码的自编码模型，都完整表达了“怎样打发时间”的意思，

表 8　联合自编码任务的正序和逆序解码生成的汉语复述实例

ID	句子来源	汉语复述实例
例子 1	Input	聪明的人如何打发他们的时间
	Reference	聪明的人怎样打发时间？
	Auto-forward	打发时间的最好方法是什么？
	Auto-Reverse	聪明的人如何打发时间？
例子 2	Input	德里的 ca-ipcc 有哪些好的培训中心？
	Reference	德里的 ca-ipcc 最好的培训机构是哪个？
	Auto-forward	德里的 ca-ipcc 最好的地方是哪里？
	Auto-Reverse	德里的 ca-ipcc 最好的培训机构是哪个？
例子 3	Input	中国和英国的餐桌礼仪有什么不同？
	Reference	中国人和英国人的餐桌礼仪有什么不同？
	Auto-forward	我们的餐桌礼仪和印度的有什么不同？
	Auto-Reverse	中国和英国的餐桌礼仪有哪些不同点？

保证了语义的一致性，验证了联合自编码任务的复述生成模型的有效性。其次，通过对

比分析联合正序解码和逆序解码自编码任务生成的复述句，我们发现联合正序解码自编码模型生成的句子并没有表达出原句中“聪明的人”的意思，而联合逆序解码自编码模型则可以准确的将较为靠前的词汇片段“聪明的人”的意思表示出来。再如例子 3，更加可以直观的感受到联合逆序自编码模型的编码器在编码靠前词汇语义信息上的有效性。

此外，我们对模型生成的汉语复述句对进行分析，发现生成的复述句可以很好的体现第 3 节的汉语复述现象。比如例子 1 中的“打发他们的时间”和“打发时间”，在不改变语义的情况下省略了某些片段，体现了第 3 节中第 11 种汉语复述现象“重复和省略互换”；例子 2 中的“培训中心”和“培训机构”，这种转换体现了第 3 节中第一种汉语复述现象的“同义词替换”。通过对生成的复述句的分析，我们可以看到基于本文构建的汉语复述平行语料，联合自编码任务的汉语复述生成模型所生成的复述句可以很好的体现汉语的复述现象，表明了本文构建的汉语复述平行语料不仅可以保证语义的一致性，而且保证了表达方式的多样性，进一步验证了本文提出的汉语复述平行语料构建方法的有效性，同时证明汉语复述现象的分类研究在汉语复述生成分析中具有重要的应用价值。

6 总结

汉语复述平行语料匮乏，严重制约了汉语复述的研究，针对这一问题，本文提出并详细介绍了基于多翻译引擎的跨语言复述知识迁移方法，构建汉语复述语料。我们设计人工评测标准，选取有道和搜狗翻译引擎翻译英语复述平行语料 Quora，构建大规模汉语复述平行语料 270k 对，单参考复述评测集 1.2k 对以及多参考复述评测集 250 对。此外，我们分析了大量汉语复述实例，结合汉语语言特征，总结汉语复述现象 13 种，其中 3 种复述现象为汉语特有，进一步说明了开展汉语复述研究具有重要意义。进一步，我们基于构建的汉语复述平行语料，设计多机制融合的神经网络复述生成模型和联合自编码任务的复述生成模型，进行汉语复述生成的评测实验。评测结果证明了提出的汉语复述平行语料构建方法可以有效构建高质量的汉语复述平行语料，用于学习汉语复述生成模型，对推动汉语复述研究提供了一定的基础数据。同时，验证了多机制融合模型和联合自编码任务模型在汉语复述生成方面的有效性。

参考文献

[1] Callisonburch C, Koehn P, Osborne M. Improved statistical machine translation using paraphrases[C]. Proc Human Language Technology Conference-north American Chapter of the Association for Computational Linguistics, 2006:17-24.

[2] 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.

[3] Sekine S. On-demand information extraction[C]. International Conference on ACL. DBLP, 2006: 731-738.

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

[5] Iordanskaja L, Kittredge R, Alain Polguère. Lexical Selection and Paraphrase in a Meaning-Text

doi:10.3772/j.issn.2095-915x.2020.05.003

Generation Model[M]. Natural Language Generation in Artificial Intelligence and Computational Linguistics. Springer US, 1991:293-312.

[6] Mckeown K R, Barzilay R, Evans D, et al. Tracking and summarizing news on a daily basis with Columbia's Newsblaster[C]. Human Language Technology Conference, 2003:280-285.

[7] Li Z, Jiang X, Shang L, et al. Paraphrase Generation with Deep Reinforcement Learning[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3865-3878.

[8] Gupta A, Agarwal A, Singh P, et al. A deep generative framework for paraphrase generation[C]. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[9] Florin Brad and Traian Rebedea. Neural paraphrase generation using transfer learning[C]. In Proceedings of the 10th Internationa Conference on Natural Language Generation, Santiago de Compostela, Spain. Association for Computational Linguistics, 2017: 257-261.

[10] 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.

[11] Zhao S, Wang H, Lan X, et al. Leveraging Multiple MT Engines for Paraphrase Generation[C]. The 23rd International Conference on Computational Linguistics Proceedings of the Main Conference (Volume 2), 2010.

[12] 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.

[13] Bhagat R, Hovy E. What is a paraphrase?[J]. Association for Computational Linguistics, 2013, 39(3): 463-472.

[14] Luong M, Le Q V, Sutskever I, et al. Multi-task Sequence to Sequence Learning[J]. arXiv preprint arXiv:1511.06114, 2015.

[15] Hasan S A, Lee K, Datla V, et al. Neural Paraphrase Generation with Stacked Residual LSTM Networks[C]. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 2923-2934.

[16] Rinaldi F, Dowdall J, Kaljurand K, et al. Exploiting Paraphrases in a Question Answering System[C]. meeting of the association for computational linguistics, 2003: 25-32.

[17] Boonthum C. iSTART: Paraphrase recognition[C]. ACL Workshop on Student Research, 2008.

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

[19] See A, Liu P J, Manning C D. Get To The Point: Summarization with Pointer-Generator Networks[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017: 1073-1083.

[20] Tu Z, Lu Z, Liu Y, et al. Modeling Coverage for Neural Machine Translation[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 76-85.

[21] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]. Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, 2004: 74-81.

[22] 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.

[23] Lavie A, Agarwal A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments[C]. Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2007: 228-231.