大规模多粒度中文复述语料库

安波 [1,2] *

（1. 中国社会科学院民族学与人类学研究所，北京，100081；2. 中国科学院软件研究所，北京，100190）

摘要：复述是相同语义的不同表达，集中反映了语言的多样性，一直是自然语言处理领域的核心问题。PPDB 英文复述数据集在英文自然语言处理的多种任务中得到了应用，推动了英文自然语言处理领域的发展。缺少大规模多粒度中文复述数据集阻碍了复述技术在中文自然语言处理中的应用，是亟待解决的问题。本文实现了一个针对多源数据的复述抽取系统，并抽取构建了一个大规模中文复述数据集，该数据集具有规模大、质量高的特点，且包含复述短语、复述模板和复述句三种粒度的复述文本。自动评估和人工评估的结果表明，我们抽取的中文复述数据具有较高的文本多样性和语义一致性。

关键词：中文复述、复述识别、复述抽取

中图分类号：H 085　　**文献标志码**：A

1 前言

复述是人类语言的一个普遍现象，美国认知心理学家 GM Olson 也将复述能力看作计算机能否理解自然语言的标准[1,2]，集中反映了语言的多样性[3,4]。如句子“北京冬奥会什么时候开幕”与“2022 年冬奥会开幕时间”表达了相似的语义，互为复述。复述在机器翻译、语义解析、问答系统和信息检索等领域具有重要应用[5,6,7]。

复述研究主要包括复述数据抽取、复述识别和复述生成三个任务，其中复述数据抽取是构建复述数据，是复述识别和复述生成模型的基础，具有更为基础性的研究价值。近年来，深度学习在自然语言处理领域得到广泛应用，基于深度学习的复述技术也被广泛地应用[8,9,10,11,12,13]。然而由于语言的多样性，基于深度学习的自然语言处理模型经常面临鲁棒性不足的问题[14]，也就是模型通常不能很好地处理相同语义的不同表达。复述通过复述识别和复述生成，可以有效地提升深度学习的鲁棒性和泛化性[7]。

PPDB[5]是被广泛应用的英文复述数据集，该数据集包括复述短语、复述模板和复述句三种不同粒度的复述数据。复述模板是指将句子或短语进一步泛化得到的抽象表示，即将句子或短语中的部分单词或短语替换为其词性表示。模板由模板词和模板槽两部分组成，模板词表示模板中具体的单词，模板槽表示模板中除单词之外的词性表示。例如，在模板“[NP1]出生于[NP2]”中，“出生于”为模板词，“[NP1]”和“[NP2]”为模板槽，复述模板在机器翻译中有重要应用[15]。除此之外，英文方面还存在其他使用较为广泛的数据集，包括 Microsoft COCO Captions 数据集[16]、PARANMT-50M 数据集[17]、PAWS 数据集[18]等。这些数据集也推动了英文机器翻译、语义解析等自然语言处理任务的发展。

目前，公开的中文复述数据集包括 PKU Paraphrase Bank 数据集[19]、BQ 数据集[20]、PAWS-X 数据集[21]、百度 Phoenix Paraphrase Dataset[1]等，这些数据集的统计信息如表 1 所示。通过表 1 可知，目前的中文复述数据集存在数据类型单一（复述句）、数据规模小等特点。缺少大规模多粒度中文复

[1] https://ai.baidu.com/broad/subordinate?dataset=paraphrasing

基金项目：国家自然科学基金（62076233）；中国社会科学院重大创新工程项目（2020YZDZX01-2）；

通讯作者：anbo724@163.com

述数据集，制约了复述技术在中文自然语言处理任务中的应用，也在一定程度上影响了基于深度学习的模型在中文自然语言处理任务中的鲁棒性和泛化性。

表 1　常用复述数据集

Tab. 1　Paraphrase Datasets

数据集	语种	句对数
MRPC	英语	4,076
QQP	英语	404K
PPDB	多语种，共 21 种语言	**416M**
PRARNMT-50M	英语	50M
PAWS	英语	108K
PAWS-X	多语种，共 6 种语言（包含中文）	320K
MS COCO Captions	英语	150M
BQ Corpus	中文	100K
PKU Paraphrase Bank	中文	509K
Phoenix Paraphrase Dataset	中文	**1M**

针对上述现状，本文实现了一种中文复述抽取方法和系统，该系统能够从多种不同类型的数据源（双语平行数据、单语可比数据和单语平行数据）中实现多种粒度（短语、模板和句子）的中文复述抽取。该系统在中英文翻译数据、电子书、电影字幕数据上实现了复述数据的抽取，得到了一个较大规模的多粒度中文复述数据集，包含复述短语、复述模板和复述句子。本文通过自动评价和人工评价的方式对抽取到的中文复述数据进行评价。实验结果表明，我们的方法抽取出的中文复述数据具有较高的语言多样性和语义一致性。

2 相关工作

本节从复述数据、复述抽取和复述识别三个方面介绍相关工作。

2.1 复述数据集

复述数据集是复述技术在自然语言处理任务中应用的基础，在英文方面已有了多种开源复述数据集，包括：PPDB 数据集[5]、PARADE 数据集[22]、Paraphrases from Twitter 数据集[23]、MS COCO Captions 数据集[16]、PARANMT-50M 数据集[17]、Diverse styles Paraphrase 数据集[24]、Opusparcus 数据集[25]和 PAWS-X 数据集[21]，以及在复述识别任务中经常使用的 MRPC 数据集、PAWS 数据集[18]、STS 数据集[26]、Quora Question Pairs 数据集[27]。其中 PPDB 和 Opusparcus 为多语种数据集。

中文复述数据集的发展较晚，目前开源的中文复述数据集包括 PAWS-X（中文）数据集、PKU Paraphrase Bank 数据集、Phoenix Paraphrase 数据集、LCQMC 数据集[28]和 BQ Corpus 数据集。以及在复述识别等评测任务中常用的数据集：CCSK2018 数据集[2]、ATEC 数据集[3]和 AFQMC 数据集[4]。从表 1 可知，无论从规模上还是类型上，中文复述数据集都还有很大的发展空间。

[2] https://www.biendata.xyz/competition/CCKS2018_3
[3] https://dc.cloud.alipay.com/index#/topic/intro?id=3
[4] https://tianchi.aliyun.com/competition

2.2 复述抽取

根据复述抽取不同的数据源，可以将复述抽取方法分为词典抽取的方法、基于双语平行语料的复述抽取方法和基于单语可比语料的复述抽取方法。

基于词典的复述抽取方法主要借助于同义词词典进行复述抽取，抽取的类型通常包含复述词（同义词）和复述句。其中复述句为同义词的不同释义。例如，从 Wordnet[29]、同义词词林[30]、大词林[31]、情感词库[32]和 Hownet[33]等语言学资源中进行复述的抽取。

基于双语平行语料的复述抽取方法以枢轴法为代表，该方法将在目标语言中具有相同翻译结果的两个源语言中的不同单词、短语或模板视为复述。该方法可以抽取复述词、复述短语和复述模板三种不同粒度的复述[34]。Ganitkevitch 等[35]利用句法解析信息从机器翻译的数据中抽取了英文词、短语和模板三种粒度的复述，并形成 PPDB 数据集，进一步地利用连续词的一致性约束来优化复述抽取的结果。李维刚等[36]通过双语短语语义约束的方法来解决短语歧义性的问题。赵世奇等[35]通过机器翻译的方法将双语平行约束转换为单语可比数据，然后进行复述的抽取。

单语可比语料包括报道同一事件的不同新闻、介绍相同事物的不同百科、对同一外文书籍的不同中文译本以及同一外文电影字幕的不同版本的翻译等，这些数据中天然地包含了大量的复述句数据，这种类型的数据被称为单语平行语料[37]。早期的研究者，利用 SVM 分类器等方法将从可比数据中抽取出来概念的不同定义作为复述句[38,39]。通过对新闻的聚类等方法，实现了从新闻数据中的复述抽取[40,41]。He 等[42]利用 tweets 中的 url 标签进行复述的抽取。近期，有研究者在 Micorsoft COCO Captions 数据集中对同一图片的不同描述作为可比语料进行复述抽取[16]。Zhang 等[19]利用相同外文著作的不同中文译本进行中文复述句抽取，并开源了复述数据集 PKU paraphrase bank。

此外，随着机器翻译的发展，有一些工作利用回译（back-translation）的方法进行复述句数据集的构建[17]。利用大规模预训练语言模型进行复述的生成也是当前研究的热点[5]。然而现有的复述生成方法本身局限于模型训练数据的语言多样性，其生成的数据的多样性也有较大局限性。通过上述从人工产生数据的复述抽取方法，更能覆盖语言的多样性，是复述抽取工作必不可少的方法。

2.3 复述识别

复述识别通过计算句子之间的语义相似度来判断给定的两个文本是否互为复述，该任务在问答系统、语义解析和信息检索等领域具有重要价值。复述识别系统也可以辅助复述抽取的过程，如在判断两个候选的句子是否为复述句等。由于复述识别的重要性和基础性，复述识别技术一直是自然语言处理领域的研究热点[2]。传统的方法包括基于特征工程和分类器进行句对的分类判断，将复述识别建模为一个二分类任务。常用的特征包括词语、句子长短、实体重叠率、编辑距离、BLEU 值等，常用的分类器包括 SVM、逻辑回归等[43,44]。一些工作还通过句法的信息来增强复述识别的准确率[35]。近年来，基于深度学习的方法在自然语言处理领域得到了广泛的应用，成为当前的主流研究方法[8,9]。目前，主流的复述识别方法也以深度学习为主，该方法将句子转换为分布式的表示，并在表示空间中计算句对是否为复述[9]。Socher 等[45]首先提出使用词向量和循环自编码器（Recursive Autoencoders）建模句法信息，然后进行句子分析，进而搭建复述识别模型。He 等[42]使用卷积神经网络来建模句子信息。Cheng 等[46]提出一种融合上下文的孪生网络（Siamese Network）的方法进行复述识别。Issa 等[47]通过句对的抽象语义表示（Abstract Meaning Representation）抽取句子的重要信息进行匹配，以达到优化复述识别的效果。针对数据不足的问题，Chen 等[57]利用强化学习来减少对训练数据的依赖。针对实体对语义的判断问题，语言知识库、知识图谱等资源被用于复述识别任务[58]。近期，随着大规模预训练语言模型的广泛使用，无监督的复述识别方法被广泛应用，如 BERT-flow[48]和 SimCSE[49]等。

3 多粒度中文复述抽取方法

本文实现了一个从双语平行语料、单语可比语料中进行大规模中文复述抽取的方法和系统，该

[5] https://github.com/Vamsi995/Paraphrase-Generator

方法能够抽取短语、模板和句子粒度的复述，形成大规模多粒度的中文复述数据集。本节主要介绍该系统的主要流程。该系统总体框架如图 1 所示，该系统针对两种不同类型的数据源分别进行了预处理和数据抽取，最终抽取了多粒度的中文复述数据。

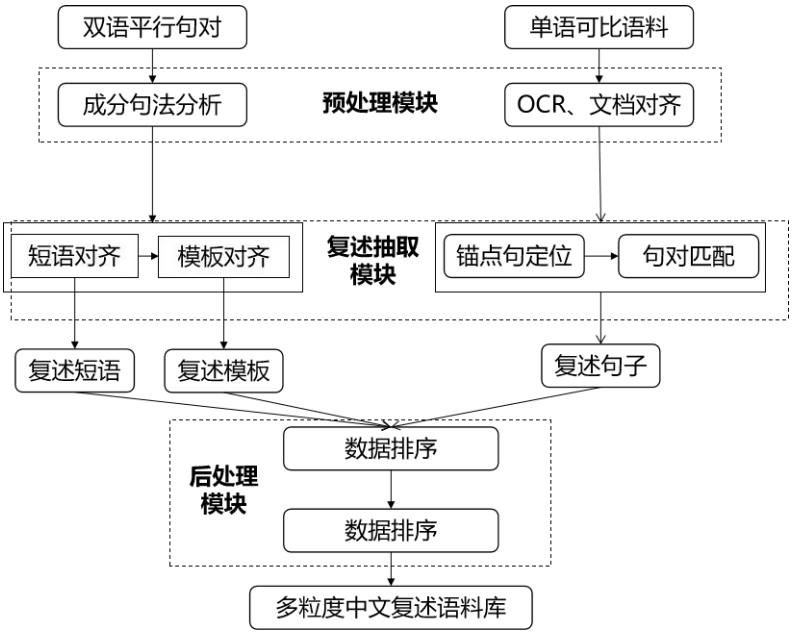

图 1 复述抽取系统总体框架

Fig. 1 The framework of paraphrase extraction system

3.1 预处理模块

不同的数据源的预处理步骤是不同的，针对双语平行语料的预处理主要包括中文分词、成分句法分析和词对齐。本文使用 Stanford CoreNLP[6]对中文数据进行分词，并对中英文句子进行成分句法分析。

词对齐是基于双语平行语料复述抽取的基础，对抽取的复述质量有重要影响。例如，PPDB 使用 GIZA++[50]进行词对齐。近年来，基于深度学习的词对齐方法被广泛使用，如 SimAlign[51]利用词向量和上下文表示，能够更好的建模词汇在不同上下文中的语义。SHIFT-AET[52]利用 Transformer[53]中的注意力机制来建模词对齐信息。本文通过集成学习方法，将三种词对齐模型的结果进行融合，得到最终的词对齐结果。具体地，本文采用加权平均的方式将三个模型输出的词对齐相似度进行集成，其计算方法如公式 1 所示。其中 $P_{Ensember}$ 为集成的词对齐概率矩阵，a_1, a_2, a_3 分别为 GIZA++、SimAlign 和 SHIFT-AET 三个词对齐模型输出的词对齐概率矩阵对应的权重，具体权重通过在验证数据上调优得到。

$$P_{Ensemble} = a_1 P_{GIZA++} + a_2 P_{SimAlign} + a_3 P_{SHIFT-ATT} \quad \text{(公式 1)}$$

本文所使用的单语可比语料主要包含电子书译本、电影字幕的不同翻译版本，这些数据保存在不同格式的图片文件中，因此需要通过 OCR 进行字符的识别，转换为文本数据。具体地，本文使用百度飞桨 PaddleOCR[7]实现字符识别。同时单语可比语料还涉及文档对齐的问题，输入的通常为一个文件集合，需要将其中的文件首先进行对齐。本文主要使用文件中的实体、时间、文件名等信息进行对齐，并在此基础上实现文件的对比。

3.2 复述抽取模块

复述抽取模块可以实现从双语平行语料中抽取复述短语、模板，从单语对比语料中抽取复述句

[6] https://stanfordnlp.github.io/CoreNLP/
[7] https://github.com/PaddlePaddle/PaddleOCR

子。下面分别介绍双语平行语料和单语可比语料的抽取过程。

3.2.1 双语平行语料复述抽取

本文复现并优化的 PPDB 复述抽取系统。对于给定的双语平行数据，通过下面步骤进行复述抽取：(1)双语句对进行词对齐；(2)基于词对齐结果抽取对齐短语；(3)从对齐短语中抽取对齐模板；(4)从对齐短语和对齐模板中抽取复述短语和复述模板。

本文中短语定义为句子语法树中的完整子树，即该子树对应的所有单词。这种类型的短语能够表达比较完整的语义，避免包含一些不相关的字、词。本文使用成分句法解析树作为短语对齐的参考。本方法包含两个步骤：短语抽取、短语对齐。

短语抽取基于成分句法解析树的结果，将成分树中的一个子树下的所有单词作为一个短语。同时，为了避免抽取出“的，你”等无意义的短语，使用子树的词性标签进行约束，仅抽取具有以下词性的短语：

英文词性约束：CD、JJ、JJR、JJS、NN、NNS、NNP、NNPS、PRP、PRP$、ADJP、NP

中文词性约束：CD、JJ、NN、NR、NT、OD、PN、ADJP、DNP、NP

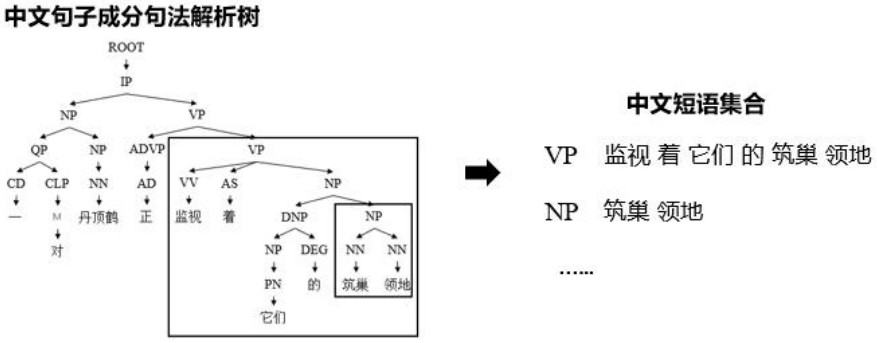

图 2 短语抽取实例

Fig. 2 Example of phrase extraction

短语对齐是找到具有语义相同关系的双语短语。Ganitkevitch[5]采用一致性约束的方法实现短语对齐，该方法要求对齐短语中所有的单词仅与被对齐的短语中的词汇对齐。该方法会导致部分短语不能很好地对齐，如图 3 所示。“丹顶鹤”应该与“red crowned cranes”对齐，但是由于单词“cranes”与单词“正”对齐，因此不满足一致性约束，无法被抽取出来。

短语对齐步骤中，PPDB 的方法采用一致性约束对短语进行对齐，即仅当两个短语中的任何一个短语都满足其中的单词仅与另一个短语中的单词对齐这一条件时，才将这两个短语对齐。我们发现一致性约束会导致部分本应对齐的短语无法对齐。例如在图 3 中，“cranes”与“正”对齐，导致“丹顶鹤”与“red crowned cranes”不满足一致性对齐条件，导致不能对齐。针对上述问题，本文通过限定词性的词汇进行一致性约束的方法来进行对齐，放松了短语对齐的条件。实验结果表明，该方法能够在引入少量噪音的情况下，显著地提升对齐短语抽取的数量。具体使用到的词性信息如下所示：

英文单词词性约束：CD、JJ、JJR、JJS、NN、NNS、NNP、NNPS、PRP、PRP$、RB、VB、VBD、VBP、VBZ

中文单词词性约束：AD、CD、NN、NR、NT、OD、VA、VV

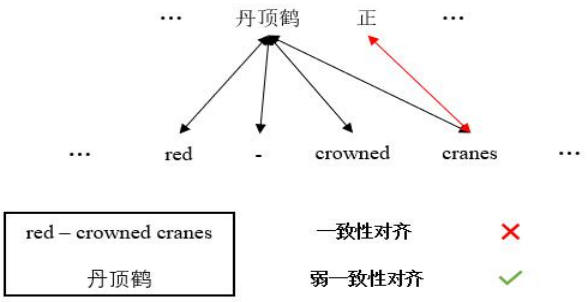

图 3　一致性约束与放松后的一致性约束

Fig. 3　consistent constraint and consistent constraint after relaxing

在抽取到对齐短语之后，通过以下方式从中抽取对齐模板（见图 4）。给定一组对齐短语，其中的部分对齐的词汇和短语使用词性进行替换后，可以形成包含部分词性信息的对齐短语，即对齐模板。根据短语中词性的个数可以分为 1 槽位和多槽位的对齐模板。

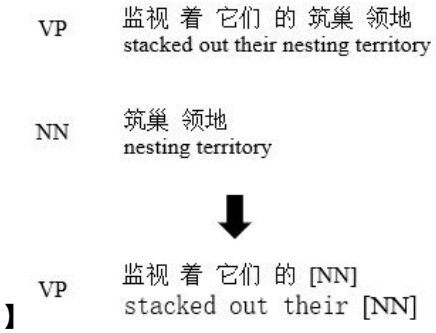

图 4　抽取对齐模板示例

Fig. 4　Example of extract aligned pattern

基于上述步骤得到的对齐短语和模板，通过找到相同英文短语/模板对应的不同的中文短语/模板即为候选中文复述短语/模板。图 5 给出了一个例子，中文短语“死于一场车祸”和“在一场车祸中丧生”均与英文短语“died in a car accident”对齐，上述两个中文短语可被抽取为候选中文复述短语。

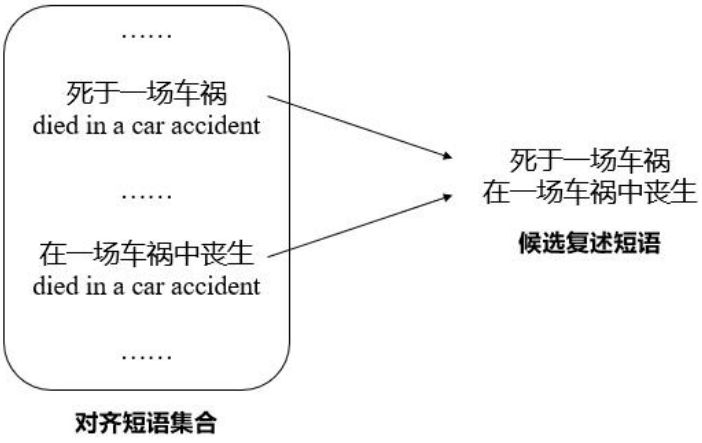

图 5　候选复述短语抽取示例

Fig. 5　Example of candidate paraphrase phrase extraction

3.2.2　单语可比语料复述抽取

单语平行语料以篇章对齐为主（如报道同一事件的新闻、同一外文名著的不同中文翻译版本），以抽取复述句为主。单语平行语料中的句子通常包含复杂的对齐关系，包括一对一、一对多和多对

多。为了抽取复述句对，本文仅选择具有一对一关系的句对进行抽取。首先，找到对齐篇章中的锚点句（显著对齐的句子），然后利用 Vecalign 算法[56]将文本分为两个部分，通过迭代上述方式进行复述句对的抽取。Vecalign 算法通过计算句子的语义距离实现句对的对齐，但传统的 Vecalign 算法依赖于字面的相似度，忽略了上下文信息对句子语义距离的影响。本文通过融合句子的上下文信息来优化句子的语义距离计算，并通过动态规划算法得到篇章中所有句对对齐的最优方案。上下文信息采用词汇对应上下文词的词向量的加权平均作为上下文信息，其中特征的权重采用 TFIDF。

3.2 后处理模块

后处理模块主要包括实体归一、特征计算、数据过滤及数据排序功能，形成最终的大规模多粒度中文复述数据集。

实体归一：不同的外文名著的中文译本中的实体名称可能采用不同的翻译，如“科诺夫尼岑”和“柯诺夫尼岑”是不同的《战争与和平》译本中人物的名称，因此需要对实体的名称进行归一化处理。

特征计算：为了提供更丰富的信息，本文借鉴 PPDB 的工作，复现并计算了其中的大部分特征，并利用 Bert 和 SentenceBert 引入了两个新的特征，（1）语言模型生成概率：基于 Bert 计算短语的生成概率；（2）基于 SentenceBert 计算两个复述文本的相似度。完整的特征信息如表 2 所示。

数据过滤：基于上述复述抽取流程，会产生一些低质量的复述，如复述对之间的文本差异很小，或者仅是实体名存在差异等情况。例如“北京到上海的高铁”和“北京到上海高铁”，仅差一个“的”，对语义的影响不大。因此需要过滤掉这部分低质量的复述数据。

数据排序：本文利用复述数据中的部分特征训练了一个回归模型，用于对候选复述数据进行排序。对复述数据进行排序后，用户可以根据对复述数据质量和数量的需要，从复述数据中选取不同规模的子集进行使用。

表 2　部分复述特征列表

Tab. 2　Some features of paraphrase data.

编号	特征	编号	特征
1	Jaccard 距离	10	source 和 target 中包含的数字是否完全相同
2	编辑距离	11	source 和 target 的词性、及词性的首字母
3	Hamming 距离	12	source 和 target 的翻译概率
4	source 是否为 target 的一部分	13	source 和 target 中是否仅包含非终结符
5	target 是否为 source 的一部分	14	source 和 target 中是否包含相邻的终结符
6	source 单词数	15	source 和 target 字符个数的差值、比值及比值的对数值
7	target 单词数	16	source 和 target 中的非终结符出现的顺序是否一致
8	source 和 target 共有的单词个数	17	source 和 target 词对齐次数
9	source 中是否存在未与 target 中的任何单词对齐的单词	18	target 中是否存在未与 source 中的任何单词对齐的单词

注：source 和 target 为具有复述关系的两个文本序列

4 中文复述抽取与数据分析

基于上述实现的中文复述数据抽取系统，本文在双语平行数据（中英文翻译数据）和单语可比

数据（电子书译本、电影字幕）上开展实验，进行复述数据的抽取。本节分别介绍复述数据评估方法、双语平行数据抽取结果、单语可比数据抽取结果。

4.1 复述数据评估指标

复述数据的评估主要包括对其多样性、语义一致性和流畅性三个方面的评估。其中，多样性表示复述句对之间表述的差异性，语义一致性表示复述句对之间语义的一致性，流畅性表示复述句对的表达是否自然、符合语法。一般情况下，仅当复述数据是采用生成的方式收集到时才需要评估其流畅性，而本文创建的复述语料库中的数据都是从自然语料中抽取出来的，因此我们不再对语料库的流畅性进行评估。本文使用自动评估和人工评估两种方式对抽取出的中文复述数据进行评估。

自动评估：本文采用复述对的编辑距离及使用长度正则化后的编辑距离作为对复述对多样性的评估指标，其中使用长度正则化后的编辑距离能够减少长度对多样性评估带来的影响。本文采用基于 SentenceBert[54]和 SimCSE 模型[49]计算出的复述数据的相似度作为语义一致性的评估指标。

人工评估：本文采用 Callison-Burch 提出的语义相似度标注方法对复述数据进行标注[55]，具体标注方法如下：当复述数据保留了原数据的所有含义，没有添加任何内容时，标注为 5 分；当复述数据保留了原数据的语义，尽管可能会添加一些附加信息，但不会改变语义时，标注为 4 分；当原数据中有些信息被删除，但不会造成太大的语义上的损失，其主要语义仍然被保留时，标注为 3 分；当复述数据与原数据的语义具有很大的差异时，标注为 2 分；当复述数据与原数据的语义完全不相关时，标注为 1 分。

4.2 双语平行语料复述抽取结果

本文使用 1000 万句对中英机器翻译数据作为数据源，开展了复述短语和复述模板的抽取，最终抽取出 239,987 对中文复述短语和 49,274,036 对中文复述模板。由于中文没有公开的复述短语和复述模板，因此我们与英文 PPDB 中的复述短语和复述模板进行对比。具体地，我们从 PPDB 数据集和抽取的数据集排序的前 20%，60%，100%的部分随机采样 500,000 条数据，然后计算其编辑距离、使用长度正则化后的编辑距离、基于 SentenceBert 模型（paraphrase-xlm-r-multilingual-v1）计算的相似度和基于 SimCSE 模型计算的相似度，其结果如表 3 和表 4 所示。其中 SimCSE 为在维基百科上随机抽取的 100w 的中英文数据分别进行训练，得到中英文的复述识别模型。

表 3　复述短语自动评估结果

Tab. 3　Automated evaluation result of paraphrase phrase

Top Ranking	数据集	编辑距离	编辑距离/长度	SentenceBert	SimCSE
20%	PPDB	2.303	**0.848**	0.584	0.781
	Ours	**5.186**	0.772	**0.917**	**0.858**
60%	PPDB	2.295	**0.904**	0.523	0.704
	Ours	**5.347**	0.893	**0.835**	**0.828**
100%	PPDB	2.317	0.921	0.504	0.698
	Ours	**5.570**	**0.955**	**0.737**	**0.713**

表 4　复述模板自动评估结果

Tab. 4　Automated evaluation result of paraphrase pattern

Top Ranking	数据集	编辑距离	编辑距离/长度	SentenceBert	SimCSE

20%	PPDB	1.682	0.447	0.838	0.77
	Ours	**3.340**	**0.775**	**0.888**	**0.783**
60%	PPDB	1.673	0.480	0.807	0.727
	Ours	**3.638**	**0.720**	**0.861**	**0.732**
100%	PPDB	1.680	0.508	0.784	0.648
	Ours	**4.07**	**0.645**	**0.781**	**0.654**

表 5　复述短语和复述模板人工评估结果

Tab. 5 Manual evaluation of paraphrase phrase and paraphrase pattern

数据类型	Top Ranking	5 分	≥4 分	≥3 分	可接受（≥3 分）
短语级复述	20%	40%	63%	87%	1,085,757
	60%	31%	52%	72%	2,695,674
	100%	23%	43%	58%	3,619,192
模板级复述	20%	52%	76%	86%	8,475,134
	60%	37%	51%	75%	22,173,316
	100%	21%	42%	61%	30,057,161

从表3可知, 本文抽取出的复述短语相较于PPDB在长度上有明显优势, 在多样性上基本与PPDB一致，在体现语义一致性的 SentenceBert 和 SimCSE 的得分上也较高。与此相对的是，在复述模板方面（表 4），本文抽取到的短语在长度、多样性方面有明显优势，在语义一致性方面与 PPDB 基本持平。

同时，我们分别从复述短语和复述模板数据的前 20%、60%、100%部分随机采样 2000 条数据进行人工标注，然后分别统计了在每一部分中，标注分数等于 5 分、大于等于 4 分、大于等于 3 分的部分所占的比例（表 5）。从标注结果可以看出，虽然在全量数据上，标注分数大于等于 3 分的百分比比较低，但是我们构建的复述数据集的规模足够大，可以根据对质量和数量的需求，选取不同规模的子集进行使用。

4.3 单语可比语料复述抽取结果

本文电子书译本和电影字幕数据作为单语可比数据开展复述抽取工作。我们基于 296 本电子书译本开展中文复述句抽取。具体地，我们自动从电子书网站进行截图，然后利用 OCR[7]技术提取文本内容，形成单语可比数据集。针对电影字幕数据，我们对 73G 的 Shooter 电影字幕合集进行了处理，在该字幕合集中主要包含两种格式的字幕文件，一种文件是由包含字幕的一组图片组成的压缩文件，针对这种类型的文件，我们仿效对电子书的处理过程，即首先利用 OCR 技术提取图片中的字幕内容，然后进行整合、分行，得到仅包含电影字幕的文本文件。另一种文件是具有特定格式的字幕文本文件，例如“.srt”格式的字幕文件，此时我们需要利用预处理模块中针对这种格式的字幕文件解析功能从中抽取出字幕内容，得到仅包含电影字幕的文本文件。

通过上述复述抽取系统，从电子书数据中共抽取出 3,097,091 对复述句对，从电影字幕数据中共抽取出 452,708 对复述句对。相比于电子书数据，尽管 Shooter 电影字幕合集包含了大量的字幕文件，但最终抽取出的复述句对却相对比较少。这是由于字幕合集中针对同一电影由不同字幕小组翻译的

不同版本的字幕比较少，而且许多字幕文件中存在大量错误信息，因此导致最终能够抽取出复述句对比较少。

同样地，针对抽取出的复述句数据，我们采用编辑距离和使用长度正则化的编辑距离作为数据多样性评估指标，采用基于 SentenceBert 和 SimCSE 语义相似度模型计算的复述句对的相似度作为语义一致性评估指标。我们利用编辑距离-语义一致性指标综合评估了复述句数据的质量（表 6）。另外，我们与 LCQMC、AFQMC、ATEC、CCKS、BQ 中文复述识别数据集中的复述数据部分以及 PKU paraphrase corpus 进行了对比。从数量上来看，本文抽取出的复述数据的规模远大于其他的数据集。因为 ATEC、BQ、CCKS 等中文复述识别数据集都经过人工筛选、标注，因此，本文的数据集与除了 LCQMC 数据集之外的小规模中文复述识别数据集相比，在数据多样性方面普遍劣于 ATEC 等中文复述识别数据集，在语义一致性方面则要优于 ATEC 等中文复述识别数据集。与 LCQMC 数据集相比，我们构建的数据集在多样性和语义一致性方面都要优于 LCQMC 数据集。与规模相对比较大的 PKU paraphrase corpus 相比，我们的数据集在数据多样性方面与之相似，在语义一致性方面则要优于 PKU paraphrase corpus。

表 6　从四种不同数据源抽取出的复述句自动评估

Fig. 6　automated evaluation of paraphrase sentence extracted from four different data source

数据集	数量	编辑距离	编辑距离/长度	SentenceBert	SimCSE
LCQMC	238,766	6.0	0.524	0.819	0.855
AFQMC	10,573	10.7	0.786	0.741	0.818
ATEC	9,158	10.0	0.751	0.752	0.828
CCKS	50,000	11.1	**0.913**	0.656	0.754
BQ	50,000	10.9	**0.913**	0.656	0.755
PKU paraphrase	509,832	**24.6**	0.658	0.797	0.854
Ours (电子书)	**3,097,091**	24.2	0.649	**0.849**	**0.861**
Ours (字幕)	452,708	10.0	0.646	0.837	0.847

注：对于本身不包含负例的数据集（如 PKU），本文通过在所有复述句子中随机替换的方法生成负例，形成复述识别数据集。

表 7　从四种不同数据源抽取出的复述句人工评估结果

Fig. 7　Manual evaluation of paraphrase extracted from four different data source

数据源	5 分	≥4 分	≥3 分	可接受（≥3 分）
电子书	46.7%	72%	95%	2,942,236
字幕	76.0%	94.3%	98.5%	445,917

同样地，我们使用人工评估的方式对复述句数据进行了评估（表 7）。其结果可知，抽取得到的复述数据绝大多数为可接受，数据质量相较复述短语和复述模板的质量更高，能够更好地推动复述技术的发展。

综上所述，本文抽取到了大规模多粒度中文复述数据集。通过与 PPDB 数据集的对比可知，我们抽取到的中文复述短语和复述模板具有较高的质量。通过与已有的中文复述数据库对比可知，我们抽取到的数据的规模更大，语义的一致性也较好。

5 总结

本文设计实现了一个大规模多粒度的中文复述抽取系统，能够从双语平行和单语可比语料中抽取多粒度的中文复述数据。本文在中英文翻译数据、电子书译本和电影字幕数据上的抽取，形成了当前最大规模的中文多粒度复述数据库 (https://github.com/casnlu/Chinese-PPDB)。自动评估和人工评估的结果表明，本系统抽取的复述短语、模板和句子具有较高的质量，能够支撑复述技术在中文领域的应用。针对中文复述应用的现状，一方面，针对中文复述句对规模较小的现状，我们计划通过挖掘更多类型的数据源来进一步增强中文复述数据集的规模。另一方面，我们计划将构建的复述数据集在复述识别、复述生成、智能问答、语义解析等任务上进行进一步的验证，并构建可以用于增强中文自然语言处理任务的复述工具集。

参考文献:

[1] Gleitman, Lila R and Gleitman, Henry. 1970. Phrase and Paraphrase: Some Innovative Uses of Language[M]. ERIC.

[2] 刘挺，李维刚，张宇，李生. 2006. 复述技术研究综述[J]. 中文信息学报. vol.20, no.4, pp.27-34.

[3] Bhagat, Rahul and Hovy, Eduard. 2003. What is a paraphrase?[J].. Computational Linguistics. vol.39, no.3, pp.463--472.

[4] De Beaugrande, Robert and Dressler, Wolfgang U. 1981. Introduction to text linguistics[M].London: longman.

[5] Ganitkevitch, Juri and Van Durme, Benjamin and Callison-Burch, Chris. 2013. PPDB: The paraphrase database[C].. NAACL , pp.758-764

[6] Bo Chen, Le Sun , Xianpei Han, Bo An. 2016. Sentence Rewriting for Semantic Parsing[C]. Meeting of the association for computational linguistics, 2016: 766-777.

[7] Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing[C], Proceedings of ACL, 2014:1415 – 1425

[8] LeCun Y, Bengio Y, and Hinton G. 2015. Deep learning[J]. Nature 521.7553(2015):436.

[9] Magnolini, S., 2014. A Survey on Paraphrase Recognition[J]. DWAI@ AI* IA, 1334, pp.33-41.

[10] El Desouki, M.I. and Gomaa, W.H., 2019. Exploring the Recent Trends of Paraphrase Detection[J]. International Journal of Computer Applications, 975, p.8887.

[11] He, H., Gimpel, K., & Lin, J. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks[C]. EMNLP, pp:1576-1586.

[12] Cheng, J., & Kartsaklis, D. (2015). Syntax-aware multisense word embeddings for deep compositional models of meaning[J]. arXiv preprint arXiv:1508.02354.

[13] Qu, Chen. , Ji, F. , Qiu, M. , Yang, L. , Min, Z. , & Chen, H. , et al. 2018. Learning to selectively transfer: reinforced transfer learning for deep text matching[J]. arXiv preprint arXiv:1812.11561

[14] Zhang, Yuan and Baldridge, Jason and He, Luheng. 2019. PAWS: Paraphrase adversaries from word scrambling[J]. arXiv:1904.01130.

[15] Muyun Yang,Junguo Zhu,Sheng Li,Tiejun Zhao.Sentence-Level Paraphrasing for Machine Translation System Combination[C]. International Conference of Young Computer Scientists.2016:612-620.

[16] Xinlei C, Hao F, Tsung-Yi L, Ramakrishna V, Saurabh G, Piotr D, C. Lawrence Z. Microsoft COCO Captions: Data Collection and Evaluation Server[J]. ArXiv, 2015 preprint arXiv:1504.00325, 2015.

[17] John W, Kevin G. PARANMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2018: 451-462.

[18] Yuan Z, Jason B, Luheng H. PAWS: Paraphrase Adversaries from Word Scrambling[C]. Proceedings of NAACL-HLT 2019. 2019: 1298-1308.

[19] Bowei Z, Weiwei S, Xiaojun W, Zongming G. PKU Paraphrase Bank : A Sentence-Level Paraphrase Corpus for Chinese[C]. In The 8th CCF International Conference on Natural Language Processing and Chinese Computing. 2019: 814-826.

[20] Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. The BQ corpus: A large-scale domain-specific Chinese corpus for sentence semantice quivalence identification[C]. In Proceedings of the 2018 EMNLP, pages 4946 – 4951, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[21] Yinfei Y, Yuan Z, Chris T, Jason B. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 3687-3692.

[22] Yun H, Zhuoer W, Yin Z, Ruihong H, James C. PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge[C]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7572-7582.

[23] Wuwei L, Siyu Q, Hua H, Wei X. A Continuously Growing Dataset of Sentential Paraphrases[C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1224-1234.

[24] Kalpesh K, John W, Mohit I. Reformulating Unsupervised Style Transfer as Paraphrase Generation[C]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 737-762.

[25] Creutz M. Open subtitles paraphrase corpus for six languages[J]. arXiv preprint arXiv:1809.06142, 2018.

[26] C. Florean, O. Bejenaru, E. Apostol, O. Ciobanu, and D. Trandabat. Sentimentalitsts at semeval-2016 task 4: building a twitter sentiment analyzer in your backyard[C]. In International Workshop on Semantic Evaluation, 2016.

[27] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs. https://www.kaggle.com/c/quora-question-pairs.

[28] Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. LCQMC:a large-scale Chinese question matching corpus[C]. In Proceedings of the 27th ACL, pages 1952 – 1962, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[29] Igor A. B, Alexander G. Synonymous Paraphrasing Using WordNet and Internet[C]. Proceedings of NLDB. 2004: 312-323.

[30] Mei, J.J., Zhu, Y.M., et al.: 1983. Tongyici Cilin[M]. Shanghai Lexicon Publishing Company, Shanghai.

[31] Wanxiang Che, Zhenghua Li, Ting Liu. 2010. LTP: A Chinese Language Technology Platformt[C]. In Proceedings of the Coling 2010: Demonstrations.08, pp13-16, Beijing, China.

[32] 陈建美, 林鸿飞, 杨志豪. 基于语法的情感词汇自动获取[J]. 2009. 智能系统学报, 4(2):100-106.

[33] Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resourcet[C]. In Proceedings of NLP-KE. IEEE, pages 820 – 824.

[34] 颜欣. 基于深度学习的细粒度复述抽取技术研究[D]. 哈尔滨工业大学, 2019.

[35] 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.

[36] 李维刚, 刘挺, 李生. 基于双语语料库的短语复述实例获取研究[J]. 中文信息学报. 2007, 21(5): 112-117.

[37] Ali I, Boris K, Jimmy L. Extracting Structural Paraphrases from Aligned Monolingual Corpora[C]. Proceedings of the Second International Workshop on Paraphrasing. 2003: 57-64.

[38] 海鹏. 基于神经网络的复述抽取和重排序研究[D]. 哈尔滨工业大学, 2015.

[39] Danni M, Chen C, Behzad G, Wang-Chiew T. Essentia: Mining Domain-specific Paraphrases with Word-Alignment Graphs[C]//Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13). 2019: 52-57.

[40] William B. D, Chris B. Automatically Constructing a Corpus of Sentential Paraphrases[C]//Proceedings of the Third International Workshop on Paraphrasing (IWP2005). 2005.

[41] Chikara H, Kentaro T, Stijn D. S, Jun’ ichi K, Sadao K. Extracting Paraphrases from Definition Sentences on the Web[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011: 1087-1097.

[42] Yun H, Zhuoer W, Yin Z, Ruihong H, James C. PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge[C]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7572-7582.

[43] Zhao SQ, Zhou M, Liu T. 2007. Learning question paraphrases for QA from Encarta logst[C]. In: Proc. of the IJCAI. Menlo Park: AAAI. 1796-1800.

[44] Brockett C, Dolan WB. 2005. Support vector machines for paraphrase identification and corpus constructiont[C]. In: Proc. of the IWP. pp:1-8.

[45] Socher, Richard, Eric H. Huang, Jeffrey Pennin, Christopher D. Manning, and Andrew Y. Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detectiont[C]. In Advances in neural information processing systems, pp. 801-809.

[46] Cheng, J., & Kartsaklis, D. (2015). Syntax-aware multisense word embeddings for deep compositional models of meaningt[J]. arXiv preprint arXiv:1508.02354.

[47] Issa, Fuad and Damonte, Marco and Cohen, Shay B and Yan, Xiaohui and Chang, Yi. 2018. Abstract meaning representation for paraphrase detectiont[C]. NAACL. pp.442-452.

[48] Li B , Zhou H , He J , et al. On the Sentence Embeddings from Pre-trained Language Models[C]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[49] Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.

[50] Franz J. O, Hermann N. A Systematic Comparison of Various Statistical Alignment Models[J]. Computational Linguistics, 2003, 29:19-51.

[51] Peter F. B, Jennifer C. L, Robert L. M. Aligning Sentences in Parallel Corpora[C]. 29th Annual Meeting of the Association for Computational Linguistics. 1991: 169-176.

[52] 周志华. 机器学习[M]. 清华大学出版社, 2016: 171-191.

[53] Jacob D, Ming-Wei C, Kenton L, Kristina T. 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). 2019: 4171-4186.

[54] Nils R, Iryna G. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 3982-3992.

[55] Zellig S. H. Distributional Structure. In The Structure of Language[J]. WORD. 1954, 10(2-3), 146-162.

[56] Brian T, Philipp K. Vecalign: Improved Sentence Alignment in Linear Time and Space[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 1342-1348.

[57] Siddique A B, Oymak S, Hristidis V. Unsupervised paraphrasing via deep reinforcement learning[C]. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1800-1809.

[58] Mohamed M, Oussalah M. A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics[J]. Language Resources and Evaluation, 2020, 54(2): 457-485.

A large Scale Multi-granularity Chinese Paraphrase Corpus

Bo An[1,2]

(1. Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing, 100081, China;
2. The Institute of Software, Chinese Academy of Sciences, Beijing, 100190, China)

Abstract: Paraphrase is the different expressions of same semantic, which reflects the diversity of languages. It has always been a core issue in the field of natural language processing. PPDB (the paraphrase database) is widely used in many natural language processing task in English, which promotes the developments of English NLP. The lack of large-scale multi-granularity Chinese paraphrase datasets hinders the application of paraphrasing technology in Chinese natural language processing, which is an urgent problem to be solved. This paper proposes and implements a Chinese paraphrase extraction system for multi-source data, and constructs a large-scale Chinese paraphrase corpus. The corpus has the characteristics of large scale and high quality, and contains paraphrase phrases, paraphrase templates and paraphrase sentences. The results of automatic evaluation and manual evaluation show that our extracted

Chinese paraphrase data has high text diversity and semantic consistency.

Keywords: Chinese Paraphrase, Paraphrase Detection, Paraphrase Extraction