文章编号：1003-0077（2011）00-0000-00

基于多译文的中文转述语料库建设及转述评价方案*

阮翀[1,2]，施文娴[1,2]，李岩昊[2]，翁伊嘉[2]，胡俊峰[1,2**]

（1.北京大学计算语言学教育部重点实验室，北京市 100871；

2. 北京大学信息科学技术学院，北京市 100871）

摘要：转述语料是转述现象研究的基础。针对目前学术界中文转述语料稀缺的现状，本文以《简爱》的多个中文译本为基础，通过句对齐得到五万句级别的平行转述语料[1]。使用无监督的小句对齐和词对齐算法，本文从语料中挖掘到九千多对词汇转述知识。同时，本文还复现和改进了机器翻译测评指标 Meteor，使得该指标更适合于中文转述句子的测评，并构造了一个中文句子转述测评数据集，以便对不同的转述知识和评价指标进行比较。实验表明，本文算法挖掘到的词汇转述知识在封闭测试中不逊于《同义词词林》。

关键词：转述知识挖掘；转述评价指标；转述语料建设

中图分类号：TP391　　**文献标识码：**A

Multi-Translation Based Chinese Paraphrase Corpus Construction and Paraphrasing Evaluation Methods

RUAN Chong[1,2], SHI Wenxian[1,2], LI Yanhao[2], WENG Yijia[2], HU Junfeng[1,2]

(1. Key Laboratory of Computational Linguistics (Ministry of Education), Peking University, Beijing, 100871, China;

2. School of Electronics and Computer Science, Peking University, Beijing, 100871, China)

Abstract: Paraphrase corpus is fundamental to research in paraphrase phenomenon, while Chinese paraphrase corpus is hardly available in academia. In this paper, we collected multiple Chinese translations of the novel Jane Eyre, and obtained around 50 thousand parallel paraphrasing sentences. Based on such corpus, we managed to extract more than 9,000 pairs of lexical paraphrase knowledge. We also reimplemented and improved a machine translation evaluation metric Meteor, making it more suitable for Chinese paraphrase evaluation. We constructed a Chinese paraphrase evaluation dataset to facilitate the comparison of different paraphrase knowledge sources and evaluation methods. The experiments show the quality of our mined knowledge is comparable to Tongyici Cilin in a closed test setting.

Key words: Paraphrase Knowledge Mining; Paraphrasing Evaluation Metric; Paraphrase Corpus Construction

---

* **收稿日期：**2018-06-10　　**定稿日期：**2018-07-25

基金项目：国家自然基金项目：大规模汉语历时语料库建设及词汇语义变迁研究（编号：61472017）。

**本文通讯作者。男，1967 年生，副教授，主要研究领域为自然语言处理。E-mail: hujf@pku.edu.cn。

[1] https://github.com/Hu-Junfeng/PKU-Chinese-Paraphrase-Corpus

1 引言

转述是指用不同的表达方式来传达相同或相似语义的语言现象。这一现象在人类语言中广泛存在，给信息检索、剽窃检测、机器翻译评价等自然语言处理问题都带来了额外的困难，而构建转述知识库可以在一定程度上缓解这一困境。

建设转述知识库需要以转述语料库为基础。目前学术界已有的转述语料多为英文的，例如 Quora[2]，MRPC[1] 和 MSCoCo[2] 等。其中 Quora 是从问答网站中收集的一些语义重复的问题对，MRPC 是从新闻语料里挖掘出的同义句对，MSCoCo 是不同人对同一图片的文字描述。而中文世界里目前还难以获得类似的公开语料，因此本文以外国文学名著的多个不同中文译本为基础，尤其是以《简爱》的四个译本为例，通过句对齐算法得到转述句对，构建了一个规模约为五万句的中文转述数据集，并在此基础上进行转述知识挖掘的相关研究。

转述现象可以在不同的层面上发生，小到词汇级别，大到篇章级别。词汇级别的转述现象最为基础和常见，一般通过同义或近义词的替换来完成。现阶段已有的中文转述知识库主要集中在词汇转述级别，本文也将重点关注从转述语料里自动提取词汇转述知识的方法。人工构建的汉语转述知识库里较为著名的是《知网》[3]和《同义词词林》[4]，尽管它们并非是为转述研究而建立的，但是其中包含的同义词汇关系使得它们成为了可用的中文转述知识库。本文将算法自动挖掘出的词汇转述知识和《同义词词林》里的转述知识进行了对比，从而验证了本文提出的词汇转述知识自动挖掘算法的有效性。

转述评价是转述研究的另一个重要组成部分，没有自动化的评价方案就难以评估转述挖掘算法的好坏、进而挖掘出更多更准确的转述知识。本文首先构建了一个转述测评数据集，然后以机器翻译中的 Meteor[5-8] 指标为基础，将转述知识引入到测评过程中，从而得到了转述知识的自动评价方案。进一步地，本文根据中文的特点引入了字符重叠知识，提出了更好的中文转述评价指标。

本文的其余部分组织结构如下：第二章介绍转述知识提取和转述评价指标的相关工作；第三章介绍本文研究所使用的语料库和转述知识提取算法，并展示分析挖掘结果；第四章介绍转述测评数据集的构建及相应测试结果。最后，第五章总结全文内容并提出未来可能的研究方向。

2 相关研究

2．1 转述知识挖掘

词汇转述知识挖掘有两大类方法，分别是从单语语料和双语平行语料中挖掘。单语语料挖掘算法总体不够成熟，常常需要依赖较为特殊的语言资源或其他复杂自然语言处理系统的辅助。例如，Wang 和 Hirst[9] 观察到字典的词条定义往往具有固定的模式，如“甲是一种乙”可以得到“甲”和“乙”具有转述关系。通过人工定义的正则表达式模板，可以提取出高质量的转述词对。而 Turney[10] 提出基于分布相似性的 PMI-IR 方法，通过使用搜索引擎检索两个候选词，统计这两个词的搜索结果的共现情况来挖掘转述词对。

基于双语平行语料的则以 Bannard 和 Callison-Burch[11] 提出的枢纽方法为代表。该方法首先收集当前语言 e 和某种枢纽语言 f 的大规模平行语料，然后训练这两种语言间的机器翻译模型，得到词汇翻译概率表，然后通过下式计算两个当前语言的单词 e_1 和 e_2 能够进行转述的概率，若概率超过一定阈值就认定转述关系成立：

$$P(e_2 \mid e_1) = \sum_f P(f \mid e_1) P(e_2 \mid f)$$

与本研究最相似的是学者 Barzilay 和 McKeown[12] 的工作，他们提出了一种自举方法从外文小说的多个英译本中提取转述词对。该方法需要训练两个分类器，一个分类器用于判定上下文是否相似，另一个分类器则用于判定中心词是否相似（是否互为转述）。其理据便是经典的分布性假设[13]：

2 https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

如果两个词相似，那么它们的上下文也相似。在算法刚启动时，首先认定相同单词出现的上下文环境是相似的，不同词出现的上下文环境则不相似，构造正负样本训练上下文分类器；然后以上下文分类器为基础，找到相似的中心词，训练优化中心词分类器。如此往复不断迭代，两个分类器都不断变优，就能挖掘到越来越多的转述词对。该算法使用词性特征来训练分类器，而小说语料中复杂多变的语言现象导致词性标注模块准确率不够高，进而产生错误累积现象。统计结果表明，算法的挖掘结果中仅有 35% 为同义词对，上下位词和兄弟词分别占 32% 和 18%，还有 11% 的词对不相关，说明该方案噪声较大。

2．2 转述评价指标

直接针对转述任务设计的评价指标很少，最有代表性的是 PEM[13]。该指标在计算时，首先需要收集当前语言和其他某种枢纽语言的大规模平行语料，然后训练两种语言间的统计机器翻译模型，得到词汇翻译概率表。对于一对当前语言的句子，可以将它们都翻译为枢纽语言的句子，通过计算翻译后句子的加权词袋相似度来给出这对句子转述程度的度量。该方法的缺点是需要收集大规模的平行语料，而且指标测评结果与训练数据有关，而不是一个清晰明了的公式。

由于转述和机器翻译具有天然的相似性，转述可以被视作单语机器翻译问题，也有很多学者直接借用机器翻译的评价指标来评测转述句子的质量，例如经典的 BLEU[14] 等指标。考虑到本研究的需求，不光需要给出句对转述质量的评价，还希望能够和转述知识库相结合，反映转述知识库本身的质量优劣。因此，本研究主要以 Meteor[5-8] 指标为基础进行改进，因为该指标在计算过程中可以引入外部转述知识。

Meteor 指标在计算时首先需要在两个句子之间寻找一个最优匹配，匹配的要求有四点，按照重要性依次递减：每个单词最多只有一个配对词、有尽可能多的单词被匹配覆盖到、最小化匹配中块的个数、最小化各匹配对之间的起始位置距离差的绝对值之和。由于上述条件可能无法同时满足，实践中通过集束搜索算法来近似找到较优解。值得一提的是，Meteor 有四种匹配模式：精确匹配、词干匹配、同义词集匹配和转述短语匹配。其中同义词集匹配和转述短语匹配需要提供额外的语言资源，从而提供了比较不同来源的转述知识库的可能。

在得到匹配结果之后，根据下式计算加权后的准确率 P 和召回率 R：

$$P = \frac{\sum_i w_i \left(\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_f)\right)}{\delta \cdot |h_c| + (1-\delta) \cdot |h_f|}$$

$$R = \frac{\sum_i w_i \left(\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f)\right)}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|}$$

其中超参数 w_i 是第 i 种类型的匹配的权重，$m_i(\cdot)$ 表示该种匹配覆盖到的词数，h 和 r 分别是指机器生成的假想译文和人工标注的参考译文，下标 c 和 f 分别是指实词和虚词（虚词定义为语料库中相对词频超过 10^{-3} 的词），超参数 δ 用于平衡实词和虚词的相对重要性。

在此之后，可以计算准确率和召回率的加权调和平均值 F_{mean}，并根据匹配中包含的块数 ch、匹配覆盖的总词数 m 得到一个句子流畅性罚分，两者相乘就是最终的 Meteor 评分（下式中 α,β,γ 均为超参数）：

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

$$Meteor = \left(1 - \gamma \left(\frac{ch}{m}\right)^{\beta}\right) \cdot F_{mean}$$

3 转述知识挖掘

本章将介绍本研究中的语料构建和处理流程：以《简爱》的四个中文译本为数据基础，首先构造句对齐语料；然后进行小句对齐和词对齐，进而得到词汇转述知识。最后展示并分析转述知识挖掘结果。

3．1 转述语料构建

本研究使用的原始生语料有些是文字版，但大多数是扫描版，然后通过 OCR 转换成文字。扫描版中时不时地会有一些文字

识别错误，例如“糟蹋”可能被识别成“糟踢”。本研究的处理流程中，首先过滤掉乱码，然后按照换行和段落信息将文本拼接和切分成句，最后使用结巴工具包[3]进行分词。分词后的一个典型例句如下：

“简，我可不喜欢吹毛求疵或者寻根究底的人；再说，小孩儿这样打断长辈的话，实在可怕。找个地方去坐下来。不会说讨人喜欢的话，就别多嘴。”

由于外文小说中常有从句嵌套的现象，导致中译本的句子长度也普遍偏长，有可能原文的一句话被拆成汉语的多个句子。因此本研究在句子划分上较为保守，划分出的句子有时是包含多个句子的一大段话，更接近于段落的概念。每个句子包含的平均单词数超过 50，更详细的统计量见下表：

表 1 《简爱》语料统计数据

《简爱》版本	0	1	2	3
句子数	4853	4952	4865	4875
词数/万词	22.6	24.8	25.0	24.2

随后本文通过微软[15]发布的 Bilingual Sentence Aligner[4] 工具包进行词对齐，其算法首先采用基于长度的方法[16]得到粗对齐结果，然后训练一个统计机器翻译模型 IBM 模型一[17]，根据这个翻译模型再筛选一遍语料，保留翻译模型认为对齐概率较大的句子。经过上述处理，《简爱》语料中共挖掘到共 24858 个句对，更详细的统计结果如下表：

表 2 《简爱》句对齐语料统计数据

版本对编号	平行句对数量
0-1	4045
0-2	3876
0-3	4090
1-2	4209
1-3	4371
2-3	4267

上表中两个句子交换顺序只计一次，因此可以通过交换句对将数据增广一倍，达到近五万对平行转述句对。其中一个转述句对示例如下（斜线表示各个小句之间的分隔）：

句子 1：“都九点了。/你是怎么搞的，/爱小姐，/让阿黛尔坐得这么久？/快带她去睡觉。”

句子 2：“九点了，/爱小姐，/你让阿黛勒坐这么久，/究竟是干什么？/带她去睡觉。”

3．2 词汇转述知识挖掘算法

以上述单语平行语料为基础，本文通过先进行小句对齐后进行词对齐的方式获取词汇转述知识，挖掘结果更加精确和全面。

本文延续 Lacoste-Julien [18]等人使用整数规划求解词对齐的思路，将对齐问题建模为如下优化问题：

$$\max_{\mathbf{0}\le\mathbf{z}\le\mathbf{1}} \sum_{jk\in E} s_{jk} z_{jk} - \sum_{\substack{j\in V^s \\ 2\le d\le D}} s_{dj\cdot} z_{dj\cdot} - \sum_{\substack{k\in V^t \\ 2\le d\le D}} s_{d\cdot k} z_{d\cdot k}$$

需要满足的两个约束条件分别为：

$$\sum_{j\in V^s} z_{jk} \le 1 + \sum_{2\le d\le D} z_{d\cdot k}, \forall k \in V^t;$$

$$\sum_{k\in V^t} z_{jk} \le 1 + \sum_{2\le d\le D} z_{dj\cdot}, \forall j \in V^s$$

其中变量 z_{jk} 表示源句子中的第 j 个词和目标句子中的第 k 个词是否匹配，s_{jk} 是匹配成功的奖励值；而变量 $z_{dj\cdot}$ 表示源语言中的第 j 个词的匹配数是否达到了 d 次，$s_{dj\cdot}$ 是对应的惩罚值，$s_{d\cdot k}$ 和 $z_{d\cdot k}$也与此类似；两个限制条件是希望每个词的总匹配次数（即 z_{jk} 之和）要符合变量 $z_{d\cdot k}$和 $z_{dj\cdot}$的要求。参数 $s_{d\cdot k}$ 和 $s_{dj\cdot}$ 应该随着 d 的增大而增大，这样才能使得模型优先选择度数较低的匹配。

原版整数规划算法只针对词对齐建模，没有考虑小句对齐的情形；还有一个重大缺陷是超参数 s 的设置需要词对齐的强监督

[3] https://pypi.org/project/jieba/

[4] https://www.microsoft.com/en-us/download/details.aspx?id=52608

数据来训练。本文则通过近年来词向量等无监督学习技术的进展直接设置超参数权重，无需训练，从而解决了词对齐标注数据缺乏的问题。具体而言，本文采用带有负采样的 word2vec 算法[19-20]训练词向量，然后根据如下公式设置单词 x 和 y 间的相似度：

$$\text{sim} = \text{cossim} + (1 - \text{cossim}) * \frac{|x \cap y|}{\max\{|x|, |y|\}}$$

其中 cossim 是两个词向量的余弦相似度，$|x|, |y|, |x \cap y|$ 分别是单词 x、单词 y、单词 x 和 y 重合部分的字符数。这种基于字符重合的修正方案可以有效增强算法的健壮性，削弱分词错误和 OCR 识别错误带来的影响。

而小句之间的相似度 s_{jk} 的设置方案为：首先枚举两个小句中的所有词对（忽略标点符号），按照上述公式计算单词相似度。如果两个小句长度都超过 5，则取其相似度排前 k=5 的词对的平均相似度为两个小句的相似度。特别地，若小句相似度超过某个阈值（本研究中取 0.95），则将小句相似度 s_{jk} 改成一个较大的数值（如 2.5），以保证整数规划算法永远选择对齐这两个小句。否则，若较短的小句长度 k<5，则取排名前 k 的相似词对的平均相似度，并按照如下方式加权得到最终的句子相似度：

$$s_{jk} = \sigma\left(\frac{\text{avg\_index}}{100} - 1\right) \sigma\left(\frac{\text{k}}{2}\right) * \text{top\_k\_sim}$$

其中 avg_index 是该小句中单词的在语料里的平均词频排名，$\sigma(\cdot)$ 是 sigmoid 函数。这两个加权项可以使得短句和常用词的权重被弱化，尤其是长度小于两个词和平均词频高于前 100 的小句会有较为显著的降权，使得算法优先考虑长句和信息量较高的小句的匹配结果。小句相似性取前几而非取平均的动机则是，两个小句里相似度最高的词对往往是真正对齐的词对，而且截断到前 5 可以更好地处理小句部分匹配和多匹配的情形。

最后，多匹配惩罚项 $s_{dj\cdot}, s_{d\cdot k}$ 的设置较为简单，只需根据词向量平均相似度和多匹配在语料中出现的频次设定一个经验值即可。本文在实验中最多允许一个小句被匹配 D=3 次，并把匹配 1 次到 3 次的惩罚值分别设定成 0.4，0.65 和 0.75。这里对单次匹配也进行惩罚的原因是，有时平行句对中的某个句子会比另一个句子多一部分内容，此时应该让这部分内容留空不做匹配，而不是强行配到某个不太合适的小句上。

上述设置已经足够处理大部分情况了，但有时会因为整数规划的多解性出现错误。例如，假设两个句子分别是“是这样！是这样”和“是的！是的！”，那么合理的匹配方式是 0-0、1-1（i-j 表示第一个句子的第 i 个小句对应第二个句子的第 j 个小句，下同），但是由于匹配 0-1、1-0 也具有同样的目标函数值，模型有可能求得这个解作为最终结果。因此，本文提出如下两趟匹配算法：

（1）第一趟先按照上述算法进行匹配，得到粗匹配结果；

（2）修正整数规划中的权重 s_{jk}。具体而言，本研究共考虑两种修正方案。其一是**对角线修正**：从粗匹配结果中找到句子 1 被匹配的第一个小句和最后一个小句的位置，分别记为 i_1 和 j_1；以及句子 2 被匹配的第一个小句和最后一个小句的位置，分别记为 i_2 和 j_2。然后对于任意一对小句 (i,j)，根据这个点到 (i_1,i_2) 和 (j_1,j_2) 的连线的距离 dist 给一个额外的奖励，奖励分值随距离指数衰减：$\text{bonus} = 0.05 * \exp(-\text{dist})$。另一种权值修正方案为**邻域强匹配修正**：如果某个位置的上下左右相邻位置有一个较为确定的匹配（小句相似度高于 0.97），就给当前位置的小句相似度加 0.1。

（3）根据修正后的小句相似度参数重新求解整数规划问题。

这种两趟匹配算法十分有效，整体匹配准确率可以达到 95%，3.1 节末尾举的复杂例子也能匹配正确，匹配结果为 0-0, 1-3, 2-1, 3-2, 4-4。

在小句对齐结果的基础上，本文进一步筛选词向量余弦相似度超过 0.75 并且共现超过两次的词对。因为小句长度较短，此时词向量余弦夹角足够小的词很可能就是互为转述的词，无需再进行词对齐步骤。

此外，本文还比较了另外两种转述词对挖掘方案，一种是将上述整数规划方法直接

用在句对齐语料上进行词对齐，跳过小句对齐的步骤；另一种是使用统计机器翻译模型在句对齐语料上寻找维特比词对齐。对于前者，只需从小句对齐算法中移除取前 k 词对相似度均值的操作，并把一对一匹配的惩罚值改成 0.3 即可。实验发现该方法准确率较高，但是召回率相对较低。而对于后者，由于统计机器翻译模型的词对齐结果不对称，本文训练两个翻译方向的词对齐模型，并通过取交集来得到更准确的结果。维特比词对齐使用 GIZA++[21]工具包得到。实验表明，当两个句子语序较为一致时，统计机器翻译模型的词对齐结果较为准确；但当语序差异较大时，往往会出现一个词对应连续多个词的情形，结果不尽如人意。

3．3 转述知识挖掘结果与分析

前一节中提到的三种转述词对挖掘方法结果汇总如下（一对词交换顺序计两次）：

表 3 三种常见转述对挖掘方法的比较

挖掘方法	规模	例词对
小句对齐+词向量过滤	8162	发表意见-说三道四 蓝-深蓝 费尽心机-煞费苦心 双眉-眉毛
整数规划词对齐	6482	有点-有点儿 说谎-撒谎 一滴-一颗 耻辱-羞辱
统计机器翻译	6322	看不清楚-不大清楚 赶忙-急急忙忙 乔治-乔治亚 外套-披风

三种方法得到的词对质量难以观察到显著的区别，但基于小句对齐和过滤具有更高的召回率。通过对三种方法的结果求并集，并人工过滤错误词对，可以得到更大规模的词汇转述知识。人工检查发现，错误类型主要是分词错误和 OCR 字符识别错误，共计不到 100 对，可见算法挖掘到的转述词对具有很高的准确率。最终合并、校验过的转述词对样例见表 4。

本文还将词汇转述关系连接拓展成网络，发现了一些有趣的子图结构，例如极大完全子图（称作转述极大团）和连通分量（转述闭包）。连通分量可以用宽度优先搜索算法来查找，而极大完全子图可以用 Bron-Kerbosch 算法[22]来枚举。经过搜索，本研究共找到 2841 个转述闭包和 5721 个转述极大团，其中一个转述闭包如图 1 所示。

表 4 《简爱》上挖掘到的常见转述对

规模	**9187**	
例词对	之内-内	山峦-群山
	疯子-疯	诧异-惊讶
	灰色-灰	预言-预见
	捆扎-捆好	努力-尽力
	争论-辩论	羞辱-耻辱
	艰难-困难	以免-免得
	宁静-安静	悔恨-后悔
	骄傲-高傲	叙述-陈述
	每天-一天	一遍-一句
	发狂-疯狂	记住-别忘了
	只须-只要	地面-地板
	屋里-进屋	寝室-宿舍

显然，表示早晨和夜晚的词不能构成转述关系，但它们却出现在了同一个转述闭包中。通过对转述极大团的分析可以发现，转述关系网络中存在“夜晚”-“今晚”-“今天”-“早上”-“早晨”这样一条路径，使得闭包中词汇的语义逐渐发生了转移。尤其是中间两个步骤：从“今晚”到“今天”发生了词义的扩大，而“今天”到“早上”又发生了词义的缩小，最终导致了词义转移现象的产生。

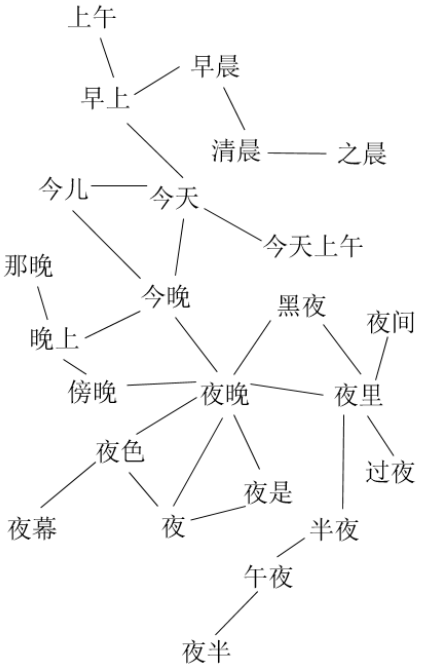

图 1 转述闭包示例

由此也能看出，转述闭包和转述极大团的语言学性质确实略有不同。转述极大团因为两两间的转述关系都得到了语料的确认，因而集合内部的联系更加紧密；而转述闭包则可能由于多次转述发生词义的扩大、缩小或偏移等现象，进而包含仅仅是话题相同但是不能互相转述的词。

4 转述知识评价

本章将对本文算法挖掘到的词汇转述知识进行测评，并和《同义词词林》等已有语言资源进行对比。同时针对中文特点，利用词汇重叠知识优化转述自动评价指标。

4．1 转述测评数据集构建

以 4 个版本的《简爱》语料为基础，本文选取各版本中能够两两对齐的句子组，随机指定其中一条语句为原句（查询语句），将剩下三条语句视为原句的转述句。然后以四版本《简爱》中的全部句子为文档集，计算每个句子和查询语句的相似度（相似度为两个句子的 TF-IDF 向量的余弦相似度乘长度惩罚项 $1-\mathrm{abs}(\mathrm{lq}-\mathrm{l})/\max\{\mathrm{lq},\mathrm{l}\}$，其中 l 和 lq 分别为候选语句和查询语句包含的词数），取相似度最高的前 5 个句子为负样本。

本研究还通过三个转述生成模型为查询语句生成三个更具迷惑性的负样本。本研究选取的基本转述生成模型是 Luong 等人[23]提出的 global attention model，唯一的区别只是将编码器部分从单向 LSTM 换成了双向。模型结构如图 2 所示，其中输入语句为 “A B C D”，输出语句为“X Y Z”（<EOS> 是用于表示句子结束的特殊符号），左侧为编码器，右侧为解码器。其他模型超参数为：编码器和解码器分别为 3 层和 2 层，LSTM 隐层和词向量的维度均为 256，词表大小为 4.5 万。该模型可以通过对目标句子的负对数似然做梯度下降来学习模型参数，即：

$$loss_{nll}=-\sum_{t=1}^{T_y}\log P(y_t\mid x,y_{<t})$$

其中 x,y 分别表示源句子和目标句子（参考转述句），T_y 是目标句子中的词数。

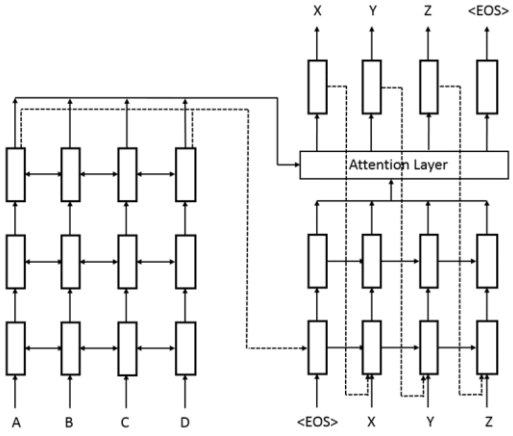

图 2 基本转述生成模型结构示意图

在基本转述生成模型之上，本研究还尝试了两种改进版模型。其一基于最近提出的词袋损失[24]，用于对不同于训练集中目标句子的正确转述句进行鼓励。该辅助损失函数认为，一个不同于参考转述句的正确转述句的词袋应该和参考转述句的词袋有较大的重合，因此只要模型生成出了参考转述句里的单词（无须考虑它是在哪一步翻译出的），就应该适当给予鼓励。其数学公式表述为：

$$loss_b(x,y) = -\sum_{t=1}^{T_y} y_t \log P_b(y_t \mid x,y)$$

$$P_b(w_i \mid x,y) = \sigma\left(\sum_{t=1}^{T_y} s_{ti}\right)$$

其中 w_i 代表词表中的任意一个词，s_{ti} 代表解码器在第 t 个时间步预测的单词 w_i 的 logits 值（未经过 softmax 归一化的概率值）。

将词袋损失和普通的负对数似然损失加权求和，便可以得到第二个转述生成模型。进一步地，可以将转述知识引入上述词袋损失中，将目标句中单词的所有转述词形成的词袋作为辅助损失计算的标准，可以对更多潜在的正确候选转述句进行鼓励，得到第三个转述生成模型：

$$loss_p(x,y) = -\sum_{t=1}^{T_y} y_t \log P_p(y_t \mid x,y)$$

$$P_p(w_i \mid x,y) = \sigma\left(\sum_{t=1}^{T_y} \sum_{(w_i,w_j)\in PP} s_{tj}\right)$$

其中 PP 是所有转述词对组成的集合。同样，该模型的总损失函数是负对数似然损失和上述转述词袋损失的加权和。

有了上述损失函数，通过梯度下降即可训练模型。三个转述生成模型的训练语料来源于《简爱》及《罪与罚》多个译本互相对齐的句对（去掉了用于构建转述测评数据集的句子），规模为接近六万个句对。所有模型均使用 Adam 算法[25]训练 10 轮；在后两个模型的训练过程中，负对数似然的权重恒为 1，而词袋损失的初始权重为 0.1，之后每一轮增加 0.1，最终增加到 1.0。

经过训练，三个模型都能生成有意义的转述句，而且迷惑性依次变强。一组具体的样例见下表中的最后一部分。

表 5 转述测评数据集示例

查询：“你要到什么地方去吗，海伦？你要回家了吗？”
真转述 1：“你要上哪儿，海伦？是回家吗？”
真转述 2：“你上哪儿去吗，海伦？你要回家是不是？”
真转述 3：“你要上哪儿去吗，海伦？你要回家去吗？”
TF-IDF 负样本 1：“可是你去的是什么地方呢，海伦？你能看到吗？你知道吗？”
转述生成模型负样本 1：“你上哪儿去呀，海伦？你已经回来了吗？”
转述生成模型负样本 2：“你上哪儿去呢，海伦？你出去没有就回来吗？”
转述生成模型负样本 3：“你要到什么地方去，海伦，你要回家了吗？”

最终构建好的转述测评数据集共包含 315 组数据，其中每组有 12 个句子：一条查询语句、三条真转述语句、五条 TF-IDF 负样本（由于空间限制，上表中只展示了其中一条）和三条转述生成模型产生的负样本。

4．2 转述测评方法

本节通过使用转述评价指标进行信息检索来比较不同转述评价指标的好坏。特别地，在转述评价指标不变的情况下，通过改变其中转述知识的来源就可以比较出转述知识的质量优劣。

具体而言，对于某种转述评价指标，本文用它计算每组测试数据中查询语句和任何一个候选语句的转述相似度，然后对结果进行排序，根据三个真转述语句出现的位置计算平均正确率均值（Mean Average Precision）。该指标越高越好。

本研究中考虑三种方案：（1）不提供转述知识，仅使用精确匹配模式计算 Meteor 指标；（2）将《同义词词林》中的底层词类作为转述知识引入 Meteor 指标中，使用精确匹配和转述匹配两种模式；（3）将本文挖掘到的转述知识加入到 Meteor 指标中，使用精确匹配和转述匹配两种模式。使用

Meteor Universal[8]中的超参数，即精确匹配和转述词匹配的权重分别为 1 和 0.6，本研究得到如下实验结果：

表 6 转述测评实验结果

转述知识源	无	《同义词词林》	本文挖掘结果
MAP	0.8453	0.8520	**0.8592**

可见效果最好的是 Meteor 加上本文挖掘到的词汇转述知识。《同义词词林》中收录词语近 7 万条，而本文挖掘到的转述词表中只有约 9 千对，却能取得更好的性能。这固然与本文进行的是封闭测试有关，但是也说明了本文算法挖掘到的转述知识库的有效性。

由于中文是孤立语，难以利用 Meteor 中针对印欧语设计的词干匹配模式。考虑到中文里相当一部分双字和多字词都符合“组合语义假设”，即词义等于字义之和，两个词有重叠的汉字往往意味着他们具有相似的语义。因此，本研究在 Meteor 的四种匹配模式外引入新的“字符重叠匹配模式”：如果组成两个单词的汉字存在重叠，就认为这两个词也能互相匹配。这种处理方式的缺点是没有分析单词的内部结构，有可能会匹配上偶然出现重合汉字的词对，例如有些汉字存在一字多义的情况。目前已有一些相关工作对汉语复合词的内部结构进行更详尽的分析，例如 CCWE[26] 使用《同义词词林》中的义类对汉语中的双字词进行标注，然后根据两个汉字的义类距离整个单词的义类的远近来学习字向量和词向量的组合关系；SCWE[27] 使用机器翻译系统将多字词内的每个字翻译成英文，然后分析每个字的翻译结果和整个词的翻译结果的相似度，据此对字向量进行自适应的加权。这些方案都有不错的效果，但模型稍显复杂。考虑到词向量也能蕴含词义信息，本文根据两个词的词向量余弦夹角进行简单的过滤，只保留词向量相似度超过一定阈值的词对，这样也能排除掉一定比例的偶然出现的汉字重叠词对。事实上，本研究也确实在实验中发现，词向量夹角校验排除掉了类似“要是-要求”这样的随机词对，提升了转述指标的效果。具体的实验结果如下（“词汇重叠匹配模式”的权重和词向量过滤阈值分别为 0.9 和 0.13，均通过网格搜索确定）：

表 7 优化后的转述测评实验结果

转述评价优化方案	字符重叠	字符重叠 + 词向量过滤
MAP	0.8954	**0.8965**

5 总结与展望

本文借助外国文学名著的多个译本构造出较大规模的中文转述平行语料，填补了目前学界这一空白。本文提出了一个健壮的、无监督的词汇转述知识提取流程，对语料中的噪声有较好的耐受能力，而且有较高的准确率和召回率。本文还构建了一个转述测评数据集，可供比较不同的转述评价指标。本文对 Meteor 指标进行了改造，使其更加适合于中文转述句子评价。

本研究以《简爱》语料的多个译本为数据基础，但是提出的算法并不依赖于具体的语料。本研究还在持续收集其他语料，如《罪与罚》等，不断补充扩大转述知识库的规模。本研究后续也将继续关注中文转述评价指标的优化工作，例如将第四章末尾提到的汉语内部构词信息考虑进来。最后，本研究还计划探索人工转述知识和算法挖掘到的转述知识相结合的方案，以及尝试把转述知识的挖掘扩展到短语级别。

6 致谢

转述语料库的建设研究得到了中央民族大学曾立英教授团队的支持和帮助。

参考文献

[1] Dolan B, Brockett C, Quirk C. Microsoft research paraphrase corpus[J]. Retrieved March, 2005, 29: 2008.

[2] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

[3] 董振东, 董强. 知网和汉语研究[J]. 当代语言学, 2001, 3(1):33-44.
[4] 梅家驹编. 同义词词林[M]. 上海辞书出版社, 1983.
[5] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005: 65-72.
[6] Denkowski M, Lavie A. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages[C]//Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Association for Computational Linguistics, 2010: 339-342.
[7] Denkowski M, Lavie A. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems[C]//Proceedings of the sixth workshop on statistical machine translation. Association for Computational Linguistics, 2011: 85-91.
[8] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language[C]//Proceedings of the ninth workshop on statistical machine translation. 2014: 376-380.
[9] Wang T, Hirst G. Exploring patterns in dictionary definitions for synonym extraction[J]. Natural Language Engineering, 2012, 18(3): 313-342.
[10] Turney P D. Mining the web for synonyms: PMI-IR versus LSA on TOEFL[C]//European Conference on Machine Learning. Springer, Berlin, Heidelberg, 2001: 491-502.
[11] Bannard C, Callison-Burch C. Paraphrasing with bilingual parallel corpora[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 597-604.
[12] Harris Z S. Distributional structure[J]. Word, 1954, 10(2-3): 146-162.
[13] Liu C, Dahlmeier D, Ng H T. PEM: A paraphrase evaluation metric exploiting parallel texts[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 923-932.
Luong T, Pham H, Manning C D. Effective Approaches to Attention
[14] 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.
[15] Moore R C. Fast and accurate sentence alignment of bilingual corpora[C]// Conference of the Association for Machine Translation in the Americas. Springer, Berlin, Heidelberg, 2002: 135-144.
[16] Gale W A, Church K W. A program for aligning sentences in bilingual corpora[J]. Computational linguistics, 1993, 19(1): 75-102.
[17] Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational linguistics, 1993, 19(2): 263-311.
[18] Lacoste-Julien S, Taskar B, Klein D, et al. Word alignment via quadratic assignment[C]//Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006: 112-119.
[19] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]//Proceedings of Workshop at ICLR, 2013.
[20] Mikolov T, Sutskever I, Chen K, et al. istributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
[21] Och F J, Ney H. A systematic comparison of various statistical alignment models[M]. MIT Press, 2003.
[22] Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph[J]. Communications of the ACM, 1973, 16(9): 575-577.
[23] Luong T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1412-1421.
[24] Ma S, Sun X, Wang Y Z, et al. Bag-of-Words as Target for Neural Machine Translation[C]//Proceedings of the ACL Conference. 2018.
[25] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.
[26] Yang L, Sun M. Improved learning of chinese word embeddings with semantic knowledge[M]//Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, Cham, 2015: 15-25.
[27] Xu J, Liu J, Zhang L, et al. Improve

chinese word embeddings by exploiting internal structure[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1041-1050.

阮翀（1993—），硕士研究生，主要研究领域为自然语言处理。
E-mail：pkurc@pku.edu.cn

施文娴（1993—），硕士研究生，主要研究领域为自然语言处理。
E-mail：wxsh@pku.edu.cn

李岩昊（1996—），本科生，主要研究领域为自然语言处理。
E-mail：eecs_lyh@pku.edu.cn