复述技术研究综述*

李维刚**，张宇，刘挺，李生
（哈尔滨工业大学计算机学院信息检索研究室，哈尔滨，150001）

摘要：复述（Paraphrase）是自然语言中比较普遍的一个现象，它集中反映了语言的多样性。复述研究的对象主要是有关短语或者句子的同义现象。随着自然语言处理各种底层技术的不断发展和成熟，对复述技术的需求也越来越强烈，因此有关复述的研究受到越来越多的关注。在英文和日文方面，复述技术已经被成功的应用到信息检索、自动问答、信息抽取、自动文摘、机器翻译以及文本水印等多个自然语言处理领域，有效地提高了这些系统的性能。本文主要对复述实例库的构建、复述规则的抽取以及复述的生成等几方面的最新研究进展进行详细的综述，并简要介绍了我们在中文复述方面进行的初步研究工作。在文章的最后一部分，我们对复述技术的难点及未来的发展方向进行了展望，并对全文进行了总结。

关键词：句子复述（改写）；复述语料库；复述抽取；复述生成；综述
中图分类号：TP391.2

Survey on Paraphrasing Technology

LI Wei-Gang, ZHANG Yu, LIU Ting, LI Sheng
(Information Retrieval Laboratory, School of Computer Science & Technology,
Harbin Institute of Technology, Harbin, 150001)

Abstract: Paraphrase is a common phenomenon in natural language which captures core aspects of variability in language. The study of paraphrase is about the synonymy phenomena of phrases or sentences. With the development of foundation technology of natural language processing, research on paraphrase has been recently received growing attention. Currently, paraphrasing technology has been applied in many NLP fields, such as, information retrieval, question answering, information extraction, automatic text summarization, machine translation and text watermark, improving the performance of these systems. This paper will mainly survey several aspects as followed: paraphrases corpus construction, paraphrases rules extraction, paraphrases generation and paraphrase evaluation. And some of our work about paraphrase is introduced in the next section. At last, some challenges, together with the future directions of paraphrasing technology are indicated.

Key words: sentence paraphrasing; paraphrases corpus; paraphrases extraction; paraphrases generation; overview

1 引言

美国认知心理学家 G.M.Olson 提出判别计算机是否理解自然语言的四条标准分别是：问答、文摘、复述和翻译。他认为，计算机只要达到了以上标准的一条，就可以说它能够理解自然语言。因此复述研究作为机器理解自然语言的标准之一，其研究价值可见一斑。随着自然语言处理各项底层技术的不断成熟和发展，复述作为自然语言中一种非常普遍的现象，受到了越来越多的研究者的关注。复述在国内也有学者称为改写[1]，其英文名称是 Paraphrase，该词的名词解释是“解释，释义等”。尽管“同义词”也属于一种广义的复述，但是在自然语言处理领域，复述研究的对象主要是“短语以上，句子以下”的语言单元，不涉及简单的词汇级同义问题，也不涉及到段落级的改写问题。目前国外对复述的研究较多，他们所处理的语言大部分是

*： 基金项目：本文受国家自然科学基金（60435020）资助
**： 作者简介：李维刚，1979 年生，男，博士研究生，研究方向为自动问答，句子复述等. Email: lee@ir.hit.edu.cn； 张宇，男，1972 年生，副教授，主要研究方向为自动问答，信息检索，文摘和抽取；刘挺，男，1972 年生，教授，主要研究方向为自然语言处理和信息检索；李生，男，1943 年生，教授，博士生导师，主要研究方向为机器翻译，自然语言处理

英语和日语，有关中文的复述研究相对较少，处于刚刚起步的阶段。本文拟从以下几个方面对复述的研究现状进行综述：首先介绍复述的定义，第三部介绍复述实例语料库的构建方法，第四部分介绍复述规则的各种抽取方法，第五部分介绍复述的生成，第六部分介绍我们在中文复述方面进行的初步工作，最后一部分对目前复述研究存在问题及发展方向进行了展望，并对全文进行了总结。

2 复述的定义

很多学者试图给复述一个精确的定义，早在二十世纪八十年代，语言学家 Halliday 和 De Beaugrande 等人就曾经给出复述的定义[5, 10]，他们认为复述保留了“概念上的近似等价”，但是互为复述的两个短语或者句子的可替换程度（Interchangeability）一直是一个没有确切标准的问题。Barzilay 等人把复述看作为传达相同信息的可替换形式，主要研究怎样抽取复述[2]。Oren Glickman 等人则认为复述现象反映了语言多变性的核心，表示对应到相同意义的等价表达[17]。

在分析了前人工作的基础上，本文主要从复述研究的对象、复述和相似度概念的区别以及复述的分类等几方面来阐述复述的概念，以更好的理解复述的内涵。

2.1 复述研究的对象

复述（Paraphrase）研究的对象主要是有关短语或者简单句的同义现象，据此，我们可以将复述研究分为两个任务：

(1) 在大规模的相关语料库中抽取短语或者简单句复述实例，构建复述实例语料库，并对复述实例库进行深层次的加工，包括词汇对齐等；

(2) 在深加工的复述实例库的基础上研究复述的生成技术，包括从中抽取复述规则并应用规则到生成上，以及利用统计的方法进行复述生成的研究等。

所谓短语是指两个或两个以上的词按照一定的语法规则构成的语法单位，简单句是指只含有一套主谓结构的句子。一般情况下，第一个任务主要是完成复述实例语料库的构建和加工，为复述的生成提供坚实的基础；第二个任务称为复述的生成，也就是将给定的一个短语或者简单句转换为另外一种形式并且保持含义不变的短语或者句子的过程。

尽管有关词汇、复杂句甚至段落也有复述现象，但是我们认为，有关短语和简单句的复述在整个复述技术本身的研究以及复述的应用研究中起着承上启下的关键作用。通过短语复述或简单句复述的分解过程可以获得同义词，通过组合过程则可以获得更复杂的复述。因此本文着重介绍有关短语和简单句级的复述研究现状。

2.2 复述和相似度概念的区别

为了更清楚的理解复述的概念，我们将复述和相似度的概念做一比较。复述的研究对象是短语和句子的同义现象，互为复述的两个短语或者句子意义是相同的，而相似度的研究对象是短语和句子的相似现象，两个相似度很高的句子可能意义完全不同。

两个句子是否为复述和相似度大小没有必然的联系。如在自动问答系统中，如果问“毛泽东的生日是哪天”，不能回答“周恩来的生日是 1 月 8 日”，尽管这两句话的相似度非常高，但他们不能称为复述；同样，用于文本水印中，也不能把文本中的“我吃了晚饭”转换为相似度很高的“我吃了早饭”。 当然，相似度也有很多的用途：在英语辅助写作系统中，如果知道“吃水果”的翻译方法，我们可以利用相似度计算知道“吃香蕉”和“吃水果”的相似度很高，因此起到了辅助写作的作用；在文本分类和聚类中，常常计算文本的相似度，把相似的文本放在一类中。

毛泽东出生于 1893 年 12 月 26 日。 （1）
伟大领袖毛泽东的生日是 1893 年 12 月 26 日。 （2）

图 1 复述实例句对

和相似度不同，复述要解决的是同义问题，对于下面的例子，由于句（2）比句（1）含有附加的信息“伟大领袖”，因此这两个句子就不能称之为复述。但是句（1）和句（2）的片断“毛泽东的生日是 1893 年 12 月 26 日”就可以组成一个复述。从上面的例子我们可以清楚的知道，对于局部同义的现象，我们必须进一步分析才能得到复述。所谓局部同义是指一个句子中的某个片段和另一个句子中的某个片段意义相同。从已有的研究来看，目前大部分的复述都是从类似的局部同义的局对中抽取出来的。

2.3 复述的分类

Bazilay等人根据复述的粒度不同，将复述分为词汇级、短语级和句子级三类。根据复述是否可以被分解，可以将复述分为原子级的复述和复合的复述[2]。而Chutima等人把常用的复述现象归纳为六类[31]，分别是：

(1) 同义词：用一个词的同义词替换该词；

(2) 语态：主动和被动语态的转换；

(3) 词性的变化：比如把一个名词用作动词，或者其它词性之间的变化；

(4) 断句：把一个长句分割成几个短句；

(5) 定义：用一个词的定义替换该词；

(6) 句子结构的变化：用不同的句子结构表达相同的意义；

Rinaldi 等人也进行过类似的划分[41]。不同类别的复述其研究方法也不同，相应的可以用在不同的应用上。由于复述现象非常普遍，因此有关复述的研究被广泛的应用到多文档文摘、自动问答系统、信息检索、信息抽取、机器翻译等领域。

2.4 复述的形式化定义

“复述”既用于表示转换一个短语或简单句从而得到其同义短语或简单句的过程（Paraphrasing），又常常用于表示该过程产生的结果（Paraphrase），本文将一个结果意义上的复述定义如下：假设两个短语或简单句 A，B，若满足以下条件：(1) A，B 为同一种语言的短语或简单句且字面不完全相同；(2) A，B 分别是结构上稳定的短语或者简单句；(3) A，B 所表达的含义相同。则称 A 为 B 的一种复述，反之亦可，称句对{A，B}为一个复述句对，简称一个复述。

特殊的，如果 A，B 分别是一个词语，则{A，B}则称为一对同义词。性质（1）主要区别于双语句对，性质（2）确定了复述研究的对象主要是“短语以上，简单句以下”的所有语言片段，而其他的比如同义词现象以及段落复述不是本文综述重点，性质（3）是{A，B}成为复述的必要条件。

3 复述实例的获取

在目前的自然语言处理研究中，并没有专门复述资源的积累，但是却存在着很多包含复述的潜在资源。比如一个外文名著对应的两个或者多个不同的译本，对同一事件的不同报道等相关的资源。对于这些文本为同一种语言，文本之间有信息的重叠，Barzilay 又将由这种文本组成的语料库称之为*相关语料库[25]。因此如何从这些相关资源中抽取复述的实例，构建复述实例语料库成为研究复述现象的一个基础工作。目前大部分有关复述的研究也都集中在此，研究的目标语言多是英语和日语。

复述实例就是不含有任何变量的原始短语或者句子复述，而复述规则是含有变量的一系列复述模板。前者可以直接从相关语料库中获取，而后者则需要一定的抽象知识，这些规则可以用到复述生成上。文献[2]将复述的获取方法大致分为有三种，第一种方法是手工获取复述，第二种方法是利用现有的语言学资源获取复述，第三种方法则是利用相关语料库获取复述。需要注意的是，有的学者将复述获取又称之为复述抽取，本文对此不加区分。本节详细介绍复述实例的获取，下一节介绍复述规则的获取。

3.1 手工获取复述

和其它资源一样，复述实例获取的第一种方法也是手工获取[12, 20]。如果仅仅获取复述实例，这种方法实现起来较为简单，但需要大量的人力物力。如果是获取复述规则，则一般需要语言学家来支持，手工获取复述规则的代价相对较大，当获取的复述规则多到一定程度，就会出现规则的冲突等弊端。这也是一种常见的知识工程的弊端，并且这种传统的手工生成的复述规则往往都是应用相关的，不易扩展，通用性也不好。目前这种获取方法只是作为辅助手段，用到一些复述获取的评价上[2, 36]。

3.2 利用现有语言学资源获取复述

利用现在已有的语言学资源也可以获取复述。针对英语可以利用的资源，例如，WordNet[16]，MindNet[19]等类似的资源。中文里类似的资源则可以利用同义词词林、知网等，但是利用语言学资源抽取出的复述大多是单词级的，根据本文规定的复述研究的对象，我们认为这种方法由于受到资源本身的限制，导致了这种方法的受限。

另外，该方法在获取复述的时候也存在着相应

*：也可称为比较语料库(comparable corpus)

的问题，比如，利用 WordNet 抽取复述，如果仅仅利用同义词作为复述，可能不够充分，这就需要一种将 WordNet 中的关系映射到复述的方法。哪些关系可以映射为复述？这在 WordNet 里并没有规定，有一些学者认为只有同义词才可以作为复述[13]，而另一些学者则放松了这种限制，规定了其他一些相似的关系也可以作为复述[27]。

利用现有的语言学资源抽取复述显然受到很多的限制，因此有的学者就通过首先构建语言学资源，然后从构建的资源中在获取复述。比如 Pereira[18], Hatzivassiloglou[11] 和 D.K. Lin[35]等人就进行了类似的工作。他们在构建资源的时候利用了分布假设（Distributional Hypothesis），其基本思想就是出现在相同上下文中的单词的意义也可能相同。这种方法虽然一定程度的扩展了资源，但是该方法抽取出的候选单词却不一定是复述，比如“香蕉”和“苹果”作为水果的不同品种经常出现在相同的上下文[35]，但是他们的意义完全不同，不能互为复述。而这种方法更为严重的一个限制就是只能抽取同义词。在后来的研究中，Lin 等人推论了分布假设，提出了扩展分布假设能够提取出句子级的推理规则[36]，这个将在后面详细介绍。另外基于语言学资源的复述方法中，还有的学者利用词典抽取复述[21]，但是也都存在着类似的问题。

3.3 基于语料库的复述获取方法

基于语料库可以抽取出多种知识资源，比如 Wu 和 Zhou 等人结合词典、双语语料库和大规模单语语料库多种资源，从中抽取同义词，进行了有意义的研究[50]。本文中基于语料库的复述获取有两层含义，第一层含义是从相关语料库中抽取复述实例，尤其是短语和句子级的复述实例[25-26, 29, 34, 42-43, 47]，又称之为复述实例的获取；第二层含义是在利用一定语言学资源的基础上，首先经过一系列的自然语言底层技术的处理，对复述实例进行一定程度的泛化和抽象，使之具有更强的表达能力，然后从中直接提取出复述规则，我们将这一部分称为复述规则的获取[30-32, 39, 41, 42, 44, 45]。本节主要介绍有关短语和句子复述实例获取的研究现状。

3.3.1 基于译文相关语料库的复述实例获取方法

很容易理解，两个作者对同一内容的不同翻译，其表达的意义一定是相同的。Barzilay 等人第一次提出利用译文相关语料构建复述语料库的方法[2]。她共搜集了三本名著的各两个译本和一本名著的三个译本（http://www.sls.csail.mit.edu/~regina/par/），利用一个经典的句子对齐技术[8]获得了译文复述语料库，共有 44,562 个复述实例句对。不过在文献中，作者提到获取句子级的复述实例只是为后续的抽取同义词和复述规则起到搜集语料库的作用。有关基于语料库进一步处理的内容，将在后面的章节里详细介绍。图 2 是一个具体的复述实例句对：

People said “The Evening Noise is sounding, the sun is setting.” “The evening bell is ringing,” people used to say.

图 2 复述实例句对

3.3.2 基于相关新闻语料复述实例获取方法

Shinyama 等人利用了一年的两份日文报纸[42]，两篇文档之间的相似度只有超过一定阈值才作为抽取复述实例的输入。其判断两个句子是否互为复述基于这样一个假设：命名实体在互为复述的句子中相同，相应的如果两个句子中含有超过一定数量的命名实体，那么这两个句子则可以组成一个复述实例。Barzilay 则利用哥伦比亚大学的 Newsblaster 系统每天在多个新闻网站上下载新闻，然后对新闻分类、聚类，并保证聚类以后获得的每一类包含的文章都是关于同一事件的[26]。从同一类的每两篇文章对中抽取复述实例句对，要求每一个句对中包含的相同单词数大于一定阈值才可以作为候选的复述。

但是利用这种方法抽取出的新闻复述实例有着先天的缺陷，最主要的一个问题就是抽取出的实例库不平衡，不能抽取出那些含有较少相同单词的复述实例。针对从新闻语料中抽取出的复述实例不平衡的问题，Bill Dolan 等人提出一种无指导抽取复述实例方法 [29]。他利用启发式策略把在同一类中新闻文章的第一个句子（可能是摘要的句子）两两组对，形成候选复述，然后设置一些过滤策略以获得最后的复述。利用这种方法可以获得一些字符相似度较低，但是意义相同的复述。

3.3.3 基于大小百科全书获取复述实例

Barzilay 等人利用大不列颠百科全书（Encyclopedia Britannica）和大不列颠基础百科全书（Britannica Elementary）作为相关语料库，从中抽取复述实例，也就是对齐单语相关语料库的问题[25]。和平行语料库对齐方法不同，文章作者采用采用的方法是利用上下文信息，结合文档的主题结

构信息学习段落匹配规则，再通过局部对齐细化匹配的段落，搜索最优的句子对作为复述。

除了上面提到的相关语料库之外，不同作者编写的同一个人的传记，医学文献上对同一种疾病的不同描述，等类似语料都是有待挖掘的非常有价值的资源，目前还没有看到利用这些资源进行复述抽取的相关研究，因此有关这方面的研究应该是非常有价值的。

4 复述规则的获取

相对来说，复述规则的获取比复述实例的获取要困难一些。加拿大多伦多大学的 Graeme Hirst[49] 对复述研究目前存在的主要问题进行了归纳，他认为下面两个问题至关重要：一个是复述知识的表示；一个是复述知识的获取。对应到复述规则获取上来，第一步就是明确规定复述规则的表示，如何表示一个规则，第二步就是怎样获取复述规则。如何将一个复述实例抽象泛化成复述规则正是这一部分研究的主要内容。

4.1 基于译文语料库的复述规则的抽取

Barzilay 利用词性序列表示复述规则。她主要利用译本相关复述实例库抽取复述规则，采用 Co-training 方法进行复述规则抽取，分以下四步：1）对齐句子中相同的单词组成初始的种子锚点；2）学习有效的上下文特征；3）学习新的复述；4）新的复述扩展锚点集合学习更多的上下文特征。所谓锚点就是确定的复述词对，这些锚点起着两个作用：第一，锚点被用来学习新的上下文规则；第二，锚点可直接用在规则里，因为这些规则包含了上下文单词的等同信息。在抽取复述规则的过程中，这种方法作为一个二元分类器，确定给定的一对短语是否是一对复述。抽取出的规则表示形式如图 3 所示：

$(NN_0\ POS\ NN_1) \leftrightarrow (NN_1\ IN\ DT\ NN_0)$ King's son　son of the king
$(IN\ NN^0) \leftrightarrow (VB^0)$ in bottles　bottled
$(VB_0\ to\ VB^1) \leftrightarrow (VB_0\ VB^1)$ start to talk　start talking
$(VB_0\ RB_1) \leftrightarrow (RB_1\ VB_0)$ suddenly came　came suddenly
$(VB\ NN^0) \leftrightarrow (VB^0)$ make appearance　appear

图 3 Co-training 方法抽取出来的规则表示形式

Ali 等人借鉴了 D.K. Lin[36]和 Barzilay[2]的方法，提出了基于句法路径抽取结构复述的方法。Barzilay 的方法利用译文语料库，抽取复述规则，要求锚点和待抽取的复述必须是相邻的。而 D.K. Lin 的方法则利用大规模的单语语料库，计算两个句子的句法结构（路径）之间的相似性，抽取出相似的路径作为复述规则，但是抽取规则的数量较少。Ali 利用译文相关语料库，通过计算复述实例中对应锚点句法路径的相似性[32]，抽取出平均长度更长的复述规则，并且能够捕捉那些长距离的依存关系，对文本进行了更深层的分析和挖掘。

4.2 基于相关新闻语料库的复述规则抽取

新闻复述语料库主要是利用同一天对同一事件的不同报导，这些不同报道中含有相同的事实信息，因此可以从中抽取出复述。Shinyama等人在信息抽取应用中抽取复述规则来支持信息抽取模式[43]，当信息抽取系统只提供一个模式的时候，其他的模式就可以根据复述推导出，从而复述能够提高信息抽取的准确率。他抽取的复述规则如图4所示：

PERSON1 shadowed *PERSON2*. *PERSON1* kept his eyes on *PERSON2*.

图4 Shinyama等人抽取出来的规则形式

其基本假设是，命名实体在不同的复述之间是相同的，比如地名，数字，人名等。该方法对复述规则的表示主要是将对应的命名实体泛化成变量抽象成信息抽取模式。

4.3 基于大规模单语语料库复述规则的抽取

Lin等人利用大规模单语语料库，计算句子的相似句法路径获取推理规则(Inference Rules)，严格来说，推理规则不完全等同于复述规则，比复述相对来说更宽泛一些[15, 36]。比如：*“X caused Y”和“Y is blamed on X”*就是一对推理规则。

Richardson曾利用路径作为特征来计算单词之间的相似度，提出一个基本假设：出现相同上下文的单词的意义也倾向于相同或相似，还有一些算法利用单词两侧固定的窗口作为他的上下文，有的算法利用依存关系作为上下文等。而Lin利用单词的相似性作为特征来计算路径的相似性，提出一个扩展的分布假设：如果两条路径倾向于连接相同的上下文，那么这两条路径的意义也倾向于相同。作者在

1G的大规模单语语料库上，在对句子进行句法分析的基础上，抽取候选的句法路径，计算路径之间的相似度，最相似的路径作为推理规则被抽取出来。抽取出的结果和TREC-QA评测用到的问题人工获取的复述进行比较，发现抽取出很多人工没有列出的丰富的复述规则。这种方法的特点是复杂度较大，获取的复述类型有限。

4.4 基于多种语料库资源的复述规则抽取

Hua Wu和Ming Zhou等人利用大规模的单语语料库结合一个有限的双语语料库，从中抽取<turn on, OBJ, light>和<switch on, OBJ, light>类似的复述规则（作者本人称之为同义搭配），解决了单独利用大规模语料库或者小规模双语语料库导致的低准确率或者低召回率的问题[52]。其具体解决办法是首先基于大规模单语语料库和一部单语词典获取候选的同义搭配，然后利用这些候选同义搭配的第二语言的译文选择最终合适的统一搭配。该方法基于这样的一个假设：如果两个搭配的译文是相似的，那么这两个搭配就是同义搭配。他们利用一个大规模语料库和一个小规模的双语语料库训练一个统计翻译模型获得候选搭配的译文。测试准确率和召回率分别为74%和64%，试验结果证明了翻译信息在选择同义搭配对时非常有效。

5 复述的生成

复述的生成就是将给定的一个短语或者句子转换为另外一个表达相同含义的短语或者句子的过程。而前面所述的复述实例语料库的构建，复述规则的抽取等都给复述的生成提供了支持。下面给出一个简单的例子，输入一个句子：“这本书多少钱？”，复述生成系统可能的一个正确输出是“这本书的价格是多少？”。复述生成作为更深层次的研究内容，成为众多研究者强烈关注的一个研究方向[3, 24, 40, 43]。其中以 Barzilay 的多重序列对齐方法[24]、Bo Pang 等人的基于句法对齐方法[3]和 Chris 等人的统计机器翻译模型方法[40]最为典型。

5.1 基于多重序列对齐的复述生成方法

Barzilay 等人提出了一种称之为多重序列对齐(Multiple Sequence Alignment，简称为 MSA)的句子级复述生成的方法[24]。首先从未标注的相关语料库中搜集结构相似的句子，多重序列对齐方法从搜集到的这些句子集合中学习一系列的由“词格子对”(Word Lattice Pair)表示的复述模式集合，然后应用这些模式生成新的复述。多重序列对齐方法主要有三步：

(1) 分别在两个语料库上计算 Lattices——以图表示——寻找一组结构相似的句子；

(2) 识别两个语料库中互为复述的 Lattices；识别过程检查两个 Lattices 中是否有相似的论元(Argument)；

(3) 给定一个待复述的句子，匹配 Lattice 并利用该 Lattice 对应部分生成复述的结果。

该方法和以往直接抽取复述的方法不同，其侧重点在生成，不仅抽取复述规则，而且自动决定潜在的相关规则，并把这些规则应用到一个新的句子上，生成形式多样的复述。另外，该方法能够生成灵活的复述类型；而只利用相关语料库和较少的知识资源。该方法的缺点是其用到的测试语料必须是与训练语料密切相关的语料。比如，在文献[24]中作者提到，训练语料是有关中东冲突的新闻语料，而测试语料是有关中东冲突的后续新闻语料，如果测试语料和训练语料是不相关的，那么这种方法的性能就会受到较大影响。

5.2 基于句法对齐的复述生成方法

Bo Pang 等人描述了一个从具有相似语义的句子集合中建立一个有限状态自动机（单词 Lattices）的基于句法对齐的复述生成算法[3]。这些有限状态（FSAs）是一种很好的复述表达方式，能够生成未见过的句子复述。作者用到的语料是在语言资源联盟（LDC）获得的，这些语料原是用来支持机器翻译评价的。有 105 个新闻片段，共 993 个句子，由 11 个不同的翻译引擎译成英文，因此每个句组都包括 11 个语义等价的译文，这些译文两两组合都可以组成复述。他们去除了含有长度大于 45 个单词的长句子的句组。该算法共分为三步：

(1) 对一个句组中的所有句子进行句法分析，把每一个句子的句法树合并到一个句法森林里；合并的过程是一个自顶向下的过程；

(2) 从句法森林里抽取有限状态自动机，也就是将句法森林映射到一个有限状态自动机的过程；采取的方法是简单的自顶向下的在句法森林的每一个合并的节点上创造一条可选择的路径，两个节点

之间的不同路径被认为是一个复述；

(3) 利用自低向上对齐的限制形式压缩句法森林，这里作者称之为挤压过程；

从理论上讲，有限状态自动机的开始节点和结束节点之间的每一条路径都对应着同一个语义表达，这样的结果就会导致可能有成千上万个句子表达同一个意思，数量非常庞大，因此里面不可避免的含有很多错误的句子。为了过滤掉这些噪声，作者在句法树合并阶段和最后的复述实例生成阶段，采用了关键词校验等一系列技术。但是，由于所用语料库的限制，并且依赖于句法分析的准确性，该方法能够生成的复述类型也是有限的。

5.3 基于统计机器翻译模型的复述生成方法

Chris等人把复述的生成过程看成是一个统计机器翻译的过程，和传统的统计机器翻译技术唯一不同的是在复述的生成过程中，源语言和目标语言是同一种语言[40]。根据统计机器翻译技术我们可知，必须有一个一定规模的复述实例语料库作为统计机器翻译的训练集。Chris用到的语料库主要是从互联网上搜集的，从相关新闻中抽取出的复述语料库[29]。

利用统计机器翻译的方法克服了许多任务相关的抽取方法的困难， Brazilay 曾提到统计机器翻译技术由于所用复述语料库的噪声以及规模的问题而不适合于复述的生成任务[2]，主要是因为，复述语料库规模太小带来了严重的数据稀疏问题。但是 Chris 提到利用大规模的语料库一定程度的解决了 Brazilay 遇到的问题。并结合了复述生成问题的自身特点，引入了短语知识库。但是，Chris 等人用到的语料库规模还远远不够大。

综上所述，目前有关复述的大部分研究仍然处于复述的获取阶段，对于真正的复述的生成还有很长的一段距离。

6 中文复述语料库的构建

尽管复述技术已经被成功的应用到多个应用领域，但是大部分都是关于英语或者日语的研究，有关中文的复述研究相对较少，处于刚刚起步的阶段。因此，为了将复述技术有效的利用到中文自然语言处理研究中来，我们对中文复述进行了针对性的初步研究。

目前主要进行了复述实例语料库的构建工作，借鉴 Barzilay 等人提出的利用名著的不同译本的方法[2]，构建复述语料库。我们利用外国名著的多个中文译本，并对多个译本进行句子对齐，获得句子级的复述实例，从而构建一个大规模的中文复述语料库。和 Barzilay 所用到的语料库不同，我们的多个平行译本都是从网络上直接得到，包含有很多噪声信息，其主要特征如下所示：

(1) 一般是篇章对齐的，但是绝大部分篇章没有严格的段落对齐；段落对齐往往是句子对齐技术的前提和基础，因此这给后续的句子对齐带来了很大的困难；

(2) 存在着翻译模式客观的复杂性，也就是说，源语言的一句话可能被翻译成两句话或多句话；

(3) 由于译者对原文意译的程度不同而存在着不同的翻译风格，导致对源语言同一句话翻译后的句子长度有着很显著的差异。

这些不规则特征的存在，导致传统对齐方法对这类文本都有着一定的局限，尤其是这些文本没有严格的段落对齐，严重影响了下一步句子对齐的精度。因此，我们提出一种基于句子长度和位置信息相结合的新方法，该方法主要基于以下假设：对齐的两个句子分布在两篇文本的相似位置，并且两个句子的长度之比倾向于等于两篇文本的整体长度之比。结合单语平行文本的特点，有效解决了没有明显段落边界文本的对齐问题，为构建大规模重述语料库提供了支持。目前，已经构建了一个 50,000 对复述实例[47]。这些复述实例将有效的支持进一步的有关复述规则的抽取以及复述的生成等研究。

7 存在的问题和展望

尽管有关复述的研究已经取得了一定的成果，但是由于语言的灵活性以及复述这种语言现象自身的特征，还存在以下问题有待深入的研究：

(1) 目前复述研究所用到的语料库绝大部分都是小规模的语料库，因此从形式各异的相关语料库中抽取规范的复述实例，从而构建一个大规模的、平衡的、知识丰富的复述语料库，将会给后续的复述研究提供一个坚实的基础；

(2) Graeme Hirst[49]也曾提过，复述规则的表示和抽取是复述研究的难点之一。目前，由于没有对句子进行深入的分析和理解，有关复述规则的抽取研究都还停留在非常简单的初始阶段，如何构建一种合理的复述规则的表达方法并自动的获取复述规则将成为复述研究的一个重大挑战；

(3) 复述实例的统计生成模型的构建。复述生成作为复述技术研究的高级阶段，将利用到各种知识资源及自然语言处理的各种底层技术，如何将多种知识资源有效的统一到一个综合生成模型中去，从而解决复述的生成问题，将是一个长远的目标；

(4) 复述的评价将有效的促进复述技术的发展，给出一个合理的评价模型能够足够好的模拟人工评价是解决评价问题的关键。对于复述的评价，目前也没有明确的标准，不过借鉴机器翻译的评价方法不失为一个有意义的探索；

(5) 复述技术的成熟离不开自然语言处理各种知识资源的支持和底层技术的成熟。具体到中文上，分词、词性标注、未登录词识别、词义消歧、句法分析以及语义分析等底层关键技术还没有完全成熟，因此怎样利用这些不完全成熟的技术促进复述的研究，并如何将复述研究的成果带动底层技术的发展将是一个值得探索的课题；

(6) 对于复杂句的复述现象，直接处理起来会比较困难。将复杂句转换为短语和简单句，首先进行短语和简单句的复述，然后通过一定的融合方式，最终获得复杂句的复述，将是一个实际可行的手段。

8 结论

本文从复述的定义，复述实例的抽取，复述规则的抽取以及复述的生成等几方面对复述技术进行了综述。我们知道，目前有关复述的大部分研究仍然处于复述的获取阶段，对于真正的复述生成技术还有很长的一段距离，处理的语言还主要是英文和日文。对于中文复述的研究，更是处于一个起步阶段，根据前文所述，有关中文复述的研究目标首先是要构建一个多种类型的大规模中文复述实例语料库，然后在此基础上，进行复述规则的获取、复述的生成以及复述的各种应用研究。相信随着自然语言处理各项底层技术的成熟，复述技术将会得到较快的发展，而复述技术的发展也必将促进自然语言底层技术的发展和成熟。最终，复述这种语言现象的本质将会被更深刻的理解。

参 考 文 献

[1] 张玉洁, 山本和英. 汉语语句自动改写. 中文信息学报, 2003, 17(6):31-38

[2] Regina Barzilay and Kathleen McKeown. Extracting paraphrases from a parallel corpus. In Proceedings of the ACL/EACL, Toulouse, 2001

[3] Bo Pang, Kevin Knight, and Daniel Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In Proceedings of HLT/NAACL, 2003

[4] P.F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 1993, 19(2):263-311

[5] De Beaugrande, Robert-Alain, and Wolfgang Dressler. Introduction to text linguistics. London; New York: Longman, 1981

[6] E. Agichtein, S. Lawrence, and L. Gravano. Learning search engine specific query transformations for question answering. In Proceedings of the 10th International World-Wide Web Conference (WWW10), 2001

[7] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1):19-52

[8] William A. Gale and Kenneth Ward Church. A program for aligning sentences in bilingual corpora. In Meeting of the Association for Computational Linguistics, 1991, pages 177-184

[9] Hal Daumé III and Daniel Marcu. Acquiring paraphrase templates from document/abstract pairs. In NL Seminar in ISI, 2003

[10] M.A.K. Halliday. An Introduction to Functional Grammar. London; Baltimore, Md, 1985.

[11] V. Hatzivassiloglou and K. R. McKeown. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In ACL 93, pages 172-182

[12] L. Iordanskaja, R. Kittredge, and A. Polguere. Lexical selection and paraphrase in a meaning-text generation model. editors, C. L. Paris, W. R. Swartout, and W. C. Mann, In Natural Language Generation in Artificial Intelligence and Computational Linguistics, Kluwer, Boston, 1991, pages 293-312

[13] Langkilde Irene and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In Proceedings of ACL/COLING, pages 704-710, 1998

[14] Beth Levin. English Verb Classes and Alternations. The University of Chicago Press, Chicago and London, 1993

[15] Dekang Lin and Patrick Pantel. DIRT-Discovery of inference rules from text. In Proceedings of the ACM

SIGKDD Conference on Knowledge Discovery and Data Mining, 2001

[16] George Miller, Richard Beckwith, Christiane Felbaum, Derek Gross, and Katherine Miller. Introduction to wordnet: An online lexical database, 1993

[17] Oren Glickman and Ido Dagan. Identifying lexical paraphrases from a single corpus: A case study for verbs. In Proceedings of Recent Advantages in Natural Language Processing, September 2003

[18] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In ACL1993, pages 183-190

[19] Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. Mindnet: Acquiring and structuring semantic information from text. Technical Report TR-98-23, Microsoft Research, 1998

[20] Jacques Robin. Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design, implementation and evaluation. PhD thesis, Columbia University, Dept of Computer Science, 1994

[21] Sadao Kurohashi and Yasuyuki Sakai. Semantic analysis of Japanese noun phrases - a new approach to dictionary-based understanding. In ACL, 1999

[22] Abdessamad Echihabi Ulf Hermjakob and Daniel Marcu. Natural language based reformulation resource and web exploitation for question answering. In Proceedings of the TREC-2002 Conference, 2002

[23] Peter Wallis. Information retrieval based on paraphrase. In Proceedings of PACLING Conference, 1993

[24] Regina Barzilay and Lillian Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In Proceedings of HLT-NAACL 2003, pages 16-23

[25] Regina Barzilay and Noemie Elhadad. Sentence alignment for monolingual comparable corpora. In Proc. of EMNLP, 2003

[26] Regina Barzilay. Information Fusion for Mutlidocument Summarization: Paraphrasing and Generation. PhD thesis, Columbia University, 2003

[27] Regina Barzilay and M. Elhadad. Using lexical chains for text summarization. In Proceedings of the ACL'97/EACL'07 Workshop on Intelligent Scalable Text Summarization, Madrid, 1997

[28] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 2002, 17:35-55

[29] Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of Coling 2004, Geneva, Switzerland, 2004, pages 350-356

[30] Florence Duclaye France. Learning paraphrases to improve a question-answering system. In EACL Natural Language Processing for Question Answering, 2003

[31] Chutima Boonthum. istart: Paraphrase recognition. In Proceedings of the Student Research Workshop: ACL. 42nd Annual Meeting of the Association of Computational Linguistics, 2004

[32] Ali Ibrahim, Boris Katz, and Jimmy Lin. Extracting structural paraphrases from aligned monolingual corpora. editors, Kentaro Inui and Ulf Hermjakob, In Proceedings of the Second International Workshop on Paraphrasing, 2003, pages 57-64

[33] Christian Jacquemin. Syntagmatic and paradigmatic representations of term variation. In 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), pages 341-348

[34] Hiroshi Kanayama. Paraphrasing rules for automatic evaluation of translation into Japanese. editors, Kentaro Inui and Ulf Hermjakob, In Proceedings of the Second International Workshop on Paraphrasing, 2003, pages 88-93

[35] Dekang Lin. Automatic retrieval and clustering of similar words. In COLING-ACL, 1998, pages 768-774

[36] Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. Natural Language Engineering, 1, 2001.

[37] F. Och and H. Ney. Improved statistical alignment models. In ACL2000, pages 440-447, Hang Kong

[38] Jahna Otterbacher and Dragomir Radev. Comparing semantically related sentences: The case of paraphrase versus subsumption. In Proceedings of Coling 2004, Geneva, Switzerland, pages 1265-1268

[39] Thierry Poibeau. Automatic extraction of paraphrastic phrases from medium-size corpora. In Proceedings of Coling 2004, Geneva, Switzerland, pages 638-644

[40] Chris Quirk, Chris Brockett, and William Dolan. Monolingual machine translation for paraphrase generation. editors, Dekang Lin and Dekai Wu, In Proceedings of EMNLP 2004, Barcelona, pages 142-149

[41] Fabio Rinaldi, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. Exploiting paraphrases in a question answering system. editors, Kentaro Inui and Ulf Hermjakob,In Proceedings of the Second International Workshop on Paraphrasing, 2003, pages 25-32

[42] Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. Automatic paraphrase acquisition from news articles, In Proceedings of Human Language Technology Conference (HLT2002), San Diego, USA, Mar. 15, 2002

[43] Yusuke Shinyama and Satoshi Sekine. Paraphrase acquisition for information extraction. editors, Kentaro Inui and Ulf Hermjakob, In Proceedings of the Second International Workshop on Paraphrasing, 2003, pages 65-71,

[44] Tetsuro Takahashi, Kozo Nawata, Shinya Kouda, Kentaro Inui, and Yuji Matsumoto. Effects of structural matching and paraphrasing in question answering. IEICE Transactions on Information and Syste, 2003.

[45] Noriko Tomuro. Interrogative reformulation patterns and acquisition of question paraphrases. editors, Kentaro Inui and Ulf Hermjakob, In Proceedings of the Second International Workshop on Paraphrasing, pages 33-40, 2003.

[46] Ingrid Zukerman, Sarah George, and Yingying Wen. Lexical paraphrasing for document retrieval and node identification. editors, Kentaro Inui and Ulf Hermjakob, In Proceedings of the Second International Workshop on Paraphrasing, 2003, pages 94-101

[47] Weigang Li, Ting Liu, Sheng Li. Combining Sentence Length with Location Information to Align Monolingual Parallel Texts. AIRS, 2004, pages 71-77

[48] Andrew FINCH, Yasuhiro AKIBA and Eiichiro SUMITA. Using a Paraphraser to Improve Machine Translation Evaluation. IJCNLP, 2004

[49] Graeme Hirst. Paraphrasing Paraphrased. In Proceedings of the Second International Workshop on Paraphrasing, 2003

[50] Hua Wu, Ming Zhou. Optimizing Synonym Extraction Using Monolingual and Bilingual Resources. In Proceedings of the Second International Workshop on Paraphrasing, 2003

[51] Mikhail J. Atallah, Victor Raskin, Mchael Crogan, Christian Hempelimann, Florian Kerschbaum, Dina Mohamed and Sanket Naik. Natural Language Watermarking: Design, Analysis, and a Proof-of-Concept Implementation. Information Hiding 2001, pages 185-199

[52] Hua Wu, Ming Zhou. Synonymous Collocation Extraction Using Translation Information. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 2003