中 文 信 息 学 报
第 20卷 第 4期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 4

[综述]

文章编号: 1003 - 0077(2006)04 - 0025 - 08

复述技术研究综述*

刘 挺,李维刚,张 宇,李 生

(哈尔滨工业大学 计算机学院信息检索研究室,黑龙江 哈尔滨 150001)

摘要:复述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短语或者句子的同义现象。自然语言处理各种底层技术的不断发展和成熟,为复述研究提高了可能,使之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、自动文摘以及机器翻译等多个领域,有效地提高了系统的性能。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:人工智能;自然语言处理;综述;句子复述;复述语料库;复述抽取;复述生成

中图分类号: TP391.2 **文献标识码**: A

A Survey on Paraphrasing Technology

LIU Ting, LI Wei-gang, ZHANG Yu, LI Sheng

(Information Retrieval Laboratory, School of Computer Science & Technology,
Haerbin Institute of Technology, Haerbin, Heilongjiang 150001, China)

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, to improve the performance of these systems. This paper will mainly survey several aspects of paraphrasing technology as followed: paraphrases corpus construction, paraphrases rules extraction, paraphrases generation and paraphrase evaluation. And some of our work about paraphrase are also introduced in brief. At the last section, some challenges, together with the future directions of paraphrasing technology are indicated.

Key words: artifical intelligence; natural language processing; overview; sentence paraphrasing; paraphrases corpus; paraphrases extraction; paraphrases generation

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

* **收稿日期**: 2005 - 08 - 09 **定稿日期**: 2006 - 06 - 05
基金项目:国家自然科学基金资助项目(60435020; 60503072; 60575042)
作者简介:刘挺(1972—),男,教授,主要研究方向为自然语言处理和信息检索.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

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。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,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 20世纪 80年代,语言学家 Halliday和 De Beaugrande等人就曾经给出复述的定义[2,3],他们认为复述保留了"概念上的近似等价",但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等人把复述看作传达相同信息的可替换形式,主要研究怎样抽取复述[4]。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[5]。

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

复述研究的对象主要是有关短语和简单句的同义现象,据此,将复述研究分为两个任务:第一,在大规模相关语料库中抽取短语或简单句的复述实例,构建复述实例语料库,并对其进行深层次的加工,包括词汇对齐等;第二,研究复述的生成技术,包括抽取复述规则应用到生成上,以及利用统计的方法进行复述生成等。

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。第一个任务主要完成复述的识别和加工;第二个任务为复述的生成。尽管有关词汇、复杂句甚至段落也有复述现象,但是有关短语和简单句的复述在其中起着承上启下的关键作用。短语或简单句复述的分解可以获得同义词,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,本文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,而相似度研究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,问题是"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。

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

复述要解决的是同义问题,对于上面的例子,由于句 (2)比句 (1)含有附加的信息"伟大领袖",因此这两个句子不能称为复述。但句 (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。从上例可知,对于局部同义的现象,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。目前大部分研究都是从类似的局部同义的句对中抽取复述。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.3 复述的分类

Bazilay等人根据复述的粒度,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[4]。而 Chutima等人把常用的复述现象归纳为六类[6],分别是:同义词、语态、词性的变化、断句、定义和句子结构的变化等。

Rinaldi等人也进行过类似的划分[7]。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。本文则把复述研究的重点分为短语和简单句两类,有关短语和简单句的复述在整个复述技术研究中起着承上启下的关键作用。通过短语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。

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,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或简单句,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[8]。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。复述实例中不含有任何变量,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。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[9,10]。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[4,11]。

3.2 利用现有语言学资源获取复述

利用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[12],MindNet[13]等类似的资源。中文里类似的资源则有同义词词林、HowNet等,这种方法抽取出的复述大多是单词级的。另外,一些学者认为只有同义词才可以作为复述[14],而一些学者则放松了这种限制,规定一些相似的关系也可以作为复述[15],目前没有一致的规定。

利用现有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,然后从中在获取复述。比如 Pereira[16], Hatzivassiloglou[17]和 D. Lin[18]和 Kurohashi[19]

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

等人就进行了类似的工作。这种方法比较严重的一个限制是只能抽取同义词。

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,从中抽取同义词资源[20]。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[8,21~26];第二层含义是在利用一定语言学资源的基础上,首先经过自然语言底层技术的处理,对实例进行一定的泛化和抽象,使之具有更强的表达能力,然后从中提取出复述规则,将这一部分称为复述规则的获取[6,7,24,27~31]。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[4]。她利用经典的句子对齐技术[32]获得了译文复述语料库。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 bellis ringing"People used to say

图 2 复述实例句对

W. Li等人也利用名著的不同译本,针对网络上直接获得的具有大量噪声的译本,提出一种基于句子长度和位置信息相结合的新方法,有效解决了没有明显段落边界文本的对齐问题,构建了一个含有约 50,000对复述实例[26]的中文复述语料库。

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[24],认为命名实体在互为复述的句子中相同,如果两个句子中含有超过一定数量的命名实体,则可以组成一个复述实例。Barzilay则在多个网站上下载新闻,然后对其分类、聚类,获得包含关于同一事件的类别[21]。从同一类的每两篇文章中抽取复述实例句对,要求句对中包含的相同单词数大于一定阈值作为候选的复述。

利用上述方法抽取复述的一个主要问题就是其不平衡性,不能抽取出那些含有较少相同单词的复述实例。针对这种不平衡问题,Dolan等人提出一种无指导抽取复述实例方法[22]。利用启发式策略把在同一类中每篇文章的第一个句子两两组对,形成候选复述,通过设置一些过滤策略以获得最后的复述。Chris Brockett[33]等人,还利用成熟的机器学习算法 SVM,在人工标注好的小部分复述实例集合上,结合多种复述特征,进一步的识别更为精确的复述实例。

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和大不列颠基础百科全书作为相关语料库,从中抽取复述实例,也就是对齐单语相关语料库的问题[8]。和平行语料库对齐方法不同,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

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Hirst[34]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[21]。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。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

Barzilay[4]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。D. Lin[11]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之间的相似性,抽取出相似的路径作为复述规则。Ali等人借鉴了两人的优点,基于句法路径从译文相关语料库中抽取出更长的复述规则,并且能够捕捉那些长距离的搭配,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[25],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。该方法直接将实例中对应的命名实体泛化成变量来表示复述规则。

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,严格来说,推理规则不完全等同于复述规则,表达更宽泛一些[11,35]。比如:"X caused Y"和"Y is blamed on X"就是一对推理规则。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:若两条路径倾向于连接相同的上下文,则这两条路径的意义也倾向于相同。利用 TREC-QA评测用到的问题,将自动抽取出的结果和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

Wu和 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,作者称之为同义搭配。解决了单独利用大规模语料库或小规模双语语料库导致的低准确率或者低召回率的问题[36]。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。

W. Li[26]等人结合一部中文语义词典和依存分析技术提出了一种基于多语义代码的复述规则的表示方法,并利用 Web上的信息直接对短语复述实例进行泛化。文章还对泛化后的复述规则进行评价,和利用词性信息表示的复述规则进行比较,取得了较好的效果。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

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或多个表达相同含义的短语或者句子的过程。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。举例说明,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,成为众多研究者十分关注的一个研究方向[25,37~39]。本文分别对基于词汇信息,句法信息和统计方法的复述生成技术进行综述。

5.1 基于词汇信息的复述生成方法

Barzilay等人提出了一种称之为多重序列对齐的句子级复述生成的方法[38]。首先从未标注的相关语料库中搜集结构相似的句子,从这些句子中学习一系列的由“词格子对”表示的复述模式集合,然后应用这些模式生成新的复述。该方法和以往直接抽取复述的方法不同,其侧重点在生成,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。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。如果测试语料和训练语料不相关,这种方法的性能就会受到较大影响。Lepage[40]等人利用一种类似基于实例的词汇或者句法变量变换的方法生成参考译文的复述实例,从而支持自动机器翻译评测,取得了较好的效果。

5.2 基于句法信息的复述生成方法

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的基于句法对齐的复述生成算法[37]。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相同语义,为此作者采用了关键词校验等一系列技术过滤其中噪声。但是,由于所用语料库的限制,该方法能够生成的复述类型也是有限的。

Stepen[41]等人利用文摘句和原文中的句子含有重合信息这一特征,从原文中抽取重要的句子并生成文摘句复述。他们利用了统计句子生成技术并结合词汇的概率生成新的句子。给定一个事件相关的句子集合,利用一个扩展版的 Veterbi算法,并采用依存关系和二元概率来发现最可能的文摘句,取得了较好的效果。

5.3 基于统计机器翻译模型的复述生成方法

Chris[39]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,源语言和目标语言是同一种语言[42]。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实例[22]。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[4],因为,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提到的问题。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 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;

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(2) Graeme Hirst[34]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初始阶段,构建一种合理的复述规则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,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,因此怎样利用现阶段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。

7 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。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,仍然处于起步阶段,根据前文所述,研究目标首先是要构建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[J]. 中文信息学报, 2003, 17(6): 31 - 38.

[2] De Beaugrande, R. Alain, and W. Dressler. Introduction to text linguistics [M]. New York: Longman, 1981.

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

[4] R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus [A]. In: ACL/EACL, 2001.

[5] O. Glickman and I. Dagan. Identifying lexical paraphrases from a single corpus: A case study for verbs [A]. In: proceedings of Recent Advantages in Natural Language Processing[A], September 2003.

[6] C. Boonthum. Istart: Paraphrase recognition [A]. In: the Student Research Workshop: ACL, 2004.

[7] F. Rinaldi, J. Dowdall, et al. Exploiting paraphrases in a question answering system [A]. In: IWP, 2003.

[8] R. Barzilay and N. Elhadad. Sentence alignment for monolingual comparable corpora [A]. In: EMNLP, 2003.

[9] L. Iordanskaja, R. Kittredge, and A. Polguere. Lexical selection and paraphrase in a meaning-text generation model [M]. In Artificial Intelligence and Computational Linguistics, 1991, pages 293 - 312.

[10] J. Robin. Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design, implementation and evaluation [D]. PhD thesis, Columbia University, 1994.

[11] D. Lin and P. Pantel. Discovery of inference rules for QA [J]. Natural Language Engineering, 1, 2001.

[12] G. Miller, R. Beckwith, et al, Introduction to wordnet: An online lexical database [M]. 1993.

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

[14] L. Irene and K. Knight. Generation that exploits corpus-based statistical knowledge [A]. In: ACL, 1998.

[15] R. Barzilay and M. Elhadad. Using lexical chains for text summarization [A]. In: ACL, 1997.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[16] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words[A]. In: ACL, 1993.
[17] V. Hatzivassiloglou and K. R. McKeown. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning [A]. In: ACL 93, pages 172 - 182.
[18] D. Lin. Automatic retrieval and clustering of similar words [A]. In: COLING-ACL, 1998, pages 768 - 774.
[19] S. Kurohashi and Y. Sakai. A new approach to dictionary-based understanding [A]. In: ACL, 1999.
[20] H. Wu, M. Zhou. Optimizing Synonym Extraction Using Mono and Bilingual Resources. [A] In: IWP [C], 2003.
[21] R. Barzilay. Information Fusion for Mutlidocument Summarization: Paraphrasing and Generation [D]. PhD thesis, Columbia University, 2003.
[22] B. Dolan, C. Quirk, et al, Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources [A]. Coling 2004.
[23] H. Kanayama. Paraphrasing rules for automatic evaluation of translation into Japanese [A]. In: IWP, 2003.
[24] Y. Shinyama, S. Sekine, et al. Automatic paraphrase acquisition from news articles [A], In: HLT, 2002.
[25] Y. Shinyama and S. Sekine. Paraphrase acquisition for information extraction [A]. In: IWP, 2003.
[26] W. Li, T. Liu. Combining Sentence Length with Location to Align Mono Parallel Texts [A]. In: AIRS, 2004.
[27] F. France. Learning paraphrases to improve a question-answering system [A]. In: EACL, 2003.
[28] A. Ibrahim, B. Katz, Extracting structural paraphrases from aligned monolingual corpora [A]. In: IWP, 2003.
[29] T. Poibeau. Automatic extraction of paraphrastic phrases from medium-size corpora [A]. In: Coling 2004.
[30] T. Takahashi, Kozo Nawata, et al. Effects of structural matching and paraphrasing in question answering [J]. IEICE Transactions on Information and Syste, 2003.
[31] N. Tomuro. Interrogative reformulation patterns and acquisition of question paraphrases [A]. In: IWP, 2003.
[32] W. Gale and K. Ward Church. A program for aligning sentences in bilingual corpora [A]. In: ACL, 1991.
[33] C. Brockett and B. Dolan. SVM for Paraphrases Identification and Corpus Construction [A]. In: IWP, 2005.
[34] G. Hirst. Paraphrasing Paraphrased [A]. In: IWP2003, 2003.
[35] D. Lin and P. Pantel. DIRT-Discovery of inference rules from text [A]. In: ACM SIGKDD, 2001.
[36] Hua Wu, Ming Zhou. Synonymous Collocation Extraction Using Translation Information[A]. In: ACL [A], 2003.
[37] BoPang, Kevin Knight, and Daniel Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences [A]. In: HLT/NAACL, 2003.
[38] R. Barzilay and L. Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment [A]. In: proceedings of HLT-NAACL 2003, pages 16 - 23.
[39] P. Brown, S. Della Pietra, et al. The mathematics of statistical machine translation: Parameter estimation [J]. Computational Linguistics, 1993.
[40] Y. Lepgae and E. Denoual. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation [A]. In: IWP, 2005.
[41] S. Wan, M. Dras. Preliminary evaluations of grammaticality [A]. In: IWP, 2005.
[42] C. Quirk, C. Brockett. Monolingual machine translation for paraphrase generation [A]. In: EMNLP 2004.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net