ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.8, August 2009, pp.2124−2137
doi: 10.3724/SP.J.1001.2009.03587
© by *Institute of Software, the Chinese Academy of Sciences*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

复述技术研究*

赵世奇+, 刘 挺, 李 生

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Research on Paraphrasing Technology

ZHAO Shi-Qi+, LIU Ting, LI Sheng

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: zhaosq@ir.hit.edu.cn

Zhao SQ, Liu T, Li S. Research on paraphrasing technology. *Journal of Software*, 2009,20(8):2124−2137. http://www.jos.org.cn/1000-9825/3587.htm

Abstract: This paper surveys the state-of-the-art research on paraphrasing in natural language processing, including the applications, the acquisition of resources, the generation, and the evaluation of paraphrases, as well as some closely related topics. This paper aims to make a summary, comparison and analysis of the mainstream methods and the latest progress in the field, expecting to be helpful to the future research.
Key words: paraphrasing; paraphrase acquisition; paraphrase generation; evaluation

摘 要: 对自然语言处理研究中的复述的研究现状与进展进行了总结,分别介绍了复述的应用、复述资源的获取、复述句的生成、复述的评测以及与复述紧密联系的相关研究等.重在对复述研究的主流方法和前沿进展进行概括、比较和分析,以期对后续研究有所助益.
关键词: 复述;复述获取;复述生成;评测
中图法分类号: TP391 **文献标识码**: A

复述(paraphrases),简单解释,就是对相同语义的不同表达[1].复述是人类语言中的一个普遍现象,由于其复杂性和多样性,前人曾对复述现象进行分类并分别予以研究.例如,Rinaldi 等人以及 Boonthum 均将常见的复述现象归纳为 6 种类型[2,3],但两者的分类结果却不尽相同.此外,我们在之前的研究中亦通过对获取的复述模板加以分析而将复述现象分为 5 类[4].本文综合已有的研究成果,将复述现象划分为以下几类:

- 细微变化.细微变化往往指的是功能词的增删与替换.例如,英文中的冠词变化(见表 1 中(a))、介词变化(见表 1 中(b)),以及中文里的量词变化等.由于这类复述引起的变化通常微小而不重要,因此应用价值不大.
- 同义短语替换.同义短语替换指的是将原文中的短语替换成其同义短语(见表 1 中(c)).其中,同义词替换可看成是同义短语替换的一个特例(见表 1 中(d)).由于这类复述在语言的实际使用中比较普遍[4]且同义短语和同义词资源相对比较容易获得,因此这类复述被广泛研究和使用.

* Supported by the National Natural Science Foundation of China under Grant Nos.60803093, 60675034 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2008AA01Z144 (国家高技术研究发展计划(863))

Received 2008-11-13; Accepted 2009-01-15

© 中国科学院软件研究所 http://www.jos.org.cn

- 词典注释替换.词典注释替换是指将原文中的词替换成词典中的注释(见表 1 中(e)).词典注释替换通常被应用于原文中的专业术语或生僻词等,通过使用比较通俗、具体的注释帮助读者更好地理解被替换词的含义.
- 语序变换.语序变换是指在不改变原文意思的前提下移动原文中某些成分的位置.在英文和中文里,时间状语、地点状语等成分的位置都比较灵活,可以出现在句首、句中或者句末,因此改变这些成分的位置通常不会改变原文的意思(见表 1 中(f)).
- 句子结构变换.句子结构变换是一种复杂的复述现象.这类复述往往不是对原文的个别成分进行简单的替换或变化,而是对原文进行较大的改动,以至于复述前后的句子结构会产生明显的变化(见表 1 中(g)).
- 句子拆分与合并.句子拆分是指将一个复杂的长句拆分成若干等价的简单短句(见表 1 中(h));句子合并则恰恰与上述过程相反.对于这类复述而言,从句的识别与指代消解等是比较关键的问题.
- 基于推理的复述.基于推理的复述是指那些需要使用某些背景知识才能正确识别和理解的复述.例如在表 1 的(i)中,只有掌握知识“首都必须是一个城市”,才能判定这两个句子意思相同.

Table 1 Examples of different types of paraphrases

表 1 各种类型的复述实例

Type	Example
Trivial change	(a) “all **the** members of” vs. “all members of” (b) “universities **of** China” vs. “universities **in** China”
Synonymous phrase substitution	(c) “I have to **take care of** the kid.” vs. “I have to **look after** the kid.” (d) “**solve** the problem” vs. “**resolve** the problem”
Dictionary definition substitution	(e) “I was suffering from **amnesia**.” vs. “I was suffering from **memory loss**.”
Phrase reordering	(f) “**Last night**, I saw Tom in the shopping mall.” vs. “I saw Tom in the shopping mall **last night**.”
Structural paraphrasing	(g) “China grows fast in economy.” vs. “China’s economic growth is fast.”
Sentence splitting and combining	(h) “He bought a computer, which is very expensive.” vs. “He bought a computer.” + “The computer is very expensive.”
Reasoning-Based paraphrasing	(i) “**Where** is the capital of China?” vs. “**What city** is the capital of China?”

复述研究可大致分为两部分,一部分是复述资源获取,包括构建复述句库、获取复述短语和复述模板等;另一部分是复述生成,即针对给定文本片段(通常是句子)生成复述.近些年来,复述作为自然语言处理(natural language processing,简称 NLP)的一个重要研究方向得到了学术界的重视,然而其研究者主要集中在欧美等国,如南加州大学、康奈尔大学、爱丁堡大学、纽约大学,以及微软研究院,Google 研究院等.上述机构主要研究英文复述,但其提出的方法通常与具体语言无关,可以比较方便地扩展到其他语言.此外,日本的东京大学、京都大学、ATR 研究院等机构在复述研究领域起步较早,研究的内容也较为广泛,但上述机构往往仅针对日文复述展开研究,其中经常涉及日文特有的语言现象和特殊处理,因此其方法的语言相关性较强.相比而言,我国的研究机构在复述方面的探索较少.中国科学院自动化研究所的宗成庆是国内较早研究复述的学者[5,6].此后,哈尔滨工业大学的李维刚等人在复述实例和模板的抽取方面进行了研究[7−9].二者的研究对象均为中文复述.除此之外,我们在之前的工作中亦在复述资源获取、复述生成以及复述应用等方面作了很多尝试[4,10−12].

1 复述的应用

由于复述现象的普遍性,自然语言处理的许多研究领域都需要解决复述的识别和生成问题,这使得复述研究有着重要的应用.其中,复述在机器翻译(machine translation,简称 MT)研究中的作用被讨论得最为广泛.前人研究表明,复述可在多个方面改进机器翻译系统的性能.首先,基于复述生成技术可以对 MT 系统的输入句进行改写,以生成更易于翻译的句子[13,14].尤其是对于口语翻译系统而言,将形式灵活且不规范的口语句子复述为规范的书面语句子无疑会降低翻译系统的处理难度[15,16].其次,复述可用于缓解统计机器翻译系统(statistical MT,简称 SMT)的数据稀疏问题.具体而言,如果待翻译的短语 e_1 在 SMT 训练语料中不存在,则 e_1 将无法被翻译.但

© 中国科学院软件研究所 http://www.jos.org.cn

倘若训练语料中存在 e_1 的复述 e_2 及 e_2 的翻译 f,则可将 f 也作为 e_1 的翻译使用,从而解决该数据稀疏问题[17].再其次,复述可用于 MT 的自动评价.MT 自动评价(如 BLEU(bilingual evaluation understudy)[18]等)的基本思想是计算机器生成译文 S 与参考译文 R 的相似度,相似度越大则 S 的分数越高.在此过程中,若可识别出 S 和 R 中那些字面表达不同但意思相同的复述片段,则可以更准确地度量二者的相似度[19−22].另外,复述也可应用于 MT 系统的参数调整[23]等方面.

自动问答系统(question answering,简称 QA)是复述应用的另一重要领域.QA 系统最关键的两个组成部分是问句理解和答案抽取,而复述在这两个部分中均有应用.在问句理解部分,可以将用户提问的一个复杂问句复述为一组简单的子句,每个子句含有原问句的一部分信息[24].如此化繁为简的过程有利于对问句的理解和后续处理.在答案抽取阶段,由于问句与答案句在表述相同信息时可能使用不同的表达方式,因此可使用复述技术来解决这一表达不匹配问题,从而提高答案抽取的召回率.这里,具体的实现方法又可分为两种.一种方法可概括为问句扩展,即将输入问句复述为多个与之等价的问句,并将这些问句全部提交给 QA 系统以抽取答案[11,25].另一种方法可概括为模板扩展,即基于复述技术生成尽可能多的答案抽取模板,以尽量涵盖答案在文档中的不同表达方式[26−28].

在信息抽取(information extraction,简称 IE)的研究中,复述也可发挥重要作用.事实上,上面介绍的答案抽取就是信息抽取在 QA 系统中的具体应用.由于基于 IE 模板的方法是信息抽取的常用方法,因此如何自动生成或扩充 IE 模板就成为信息抽取研究的关键问题.研究证明,使用复述技术可以自动生成大量丰富的 IE 模板,从而提高 IE 系统的性能[29−31].此外,在信息抽取的子领域——关系抽取(relation extraction,简称 RE)的研究中,复述技术亦被证明可用于 RE 模板的自动生成[32,33].

复述在信息检索(information retrieval,简称 IR)中的应用主要体现为查询扩展.由于 IR 系统的输入查询通常是关键词的序列,因此其中使用的一般是词汇级的复述.已有研究表明,基于复述的查询扩展可以有效地提高 IR 系统的性能[34].

在自动文摘(summarization)的研究中,复述的作用体现在两个方面.一方面,复述可用于计算句子相似度,识别出那些意思相同或相近的句子.这样可以更好地进行句子聚类、文摘句选择,并生成更加准确、凝练的文摘[35].另一方面,复述还可用于机器生成文摘的自动评价.与前面介绍的机器翻译的自动评价一样,通过正确识别机器生成文摘与参考文摘之间的复述,可以更准确地为生成文摘打分,进而有助于对文摘系统性能的评价[36].

在自然语言生成(natural language generation,简称 NLG)的研究中,复述可被应用于自动生成句子的改写,以使生成的句子更加生动、通顺和富于变化.尤其是在 NLG 的词汇选择环节,在表达一个特定语义时,如何根据不同的上下文灵活地选择和变换所使用的词汇,事实上就属于复述的研究范畴[37].

除了上面介绍的几个主要应用领域以外,复述在其他一些领域也扮演着重要的角色.例如,复述技术可用于句子的自动校正,即把存在用词或句法错误的句子复述为正确的句子[38].复述还可用于辅助阅读,通过将复杂、深奥的文本复述为简单易懂的文本,以减轻有语言障碍者的阅读困难[39].在密码学和文本水印的研究中,复述可用于句子改写,进而达到信息隐藏的目的[40].另外,复述技术还有助于对文章抄袭和剽窃现象的自动发现,即在抄袭者对原文的某些内容进行了复述的情况下,仍可将其识别出来[41].总之,复述在以上众多研究领域的应用,使其成为一个非常重要的研究方向.

2 复述资源获取

复述资源获取即利用词典、语料库以及互联网等获取复述资源.这里,复述资源包括粒度较大的复述句,也包括粒度较小的复述短语和复述模板.其中,获取的复述句通常被用于进一步抽取复述短语和模板,或者直接用作基于统计的复述生成的训练语料;而复述短语和复述模板则可直接应用到前面介绍的机器翻译、自动问答、信息抽取等各种研究中.

2.1 复述句的获取

在统计自然语言处理研究中,训练语料是统计学习得以进行的前提.例如,在统计机器翻译中,一个大规模

© 中国科学院软件研究所 http://www.jos.org.cn

的双语平行语料库是必不可少的.因此,在复述研究中,人们自然想到可以首先构建一个含有大量复述句对的语料库,然后再从中统计出各种复述现象.以此为出发点,许多研究者探索如何获取大量的复述句并构建复述句库.人们首先想到的是现实世界中已经存在的复述句对,即对同一个外文句子的不同英文翻译.在文学领域,一部好的外语文学作品经常在不同时期被不同人翻译成多个译本.这些译本往往忠实于原文,即章节段落的划分、句意等都会与原文保持一致.但由于个人习惯和写作风格等方面的差异性,各译本在具体表达上均不相同,故而可以使用句对齐技术从这些平行译本中抽取复述句[1,42].另外,也有人利用机器翻译评测的参考答案构建复述句库[43].例如,在NIST(National Institute of Standards and Technology)机器翻译自动评测**中,每个测试句会提供多个参考译文,这些译文即互为复述.上述方法最大的优点在于简单易行,缺点则在于获取的复述句对数量较少且领域局限性较大.比如,著作多译本往往仅存在于文学领域,对于其他领域及其他文体则很难找到.机器翻译的参考译文更是数量有限且增长缓慢.

另外一类用于抽取复述句的资源是可比(comparable)新闻语料[29,44−46].所谓可比新闻是指报道同一事件的多篇新闻.这些新闻主题相同,内容近似,其中可能存在部分复述句.随着互联网的普及与发展,多家网站同时报道某一突发事件或热点新闻的现象很常见,因此互联网上存在大量可比新闻.然而,这里有两个问题有待解决,一是如何自动收集可比新闻,二是如何从可比新闻中进一步抽取出复述句.对于前一个问题,研究者们一般采用聚类的方法实现可比新闻的自动发现,而聚类算法主要考虑的是新闻内容的相似度和新闻发布时间的间隔等[44,45].对于后一个问题,研究者们提出了多种复述句抽取方法.比如,有人根据总结新闻的写作习惯,认为每篇新闻的前两句一般为全篇的摘要,而各篇可比新闻的摘要则应意思相同.因此,他们提出一种简单策略,即把可比新闻的前两句提取出来相互配对而构成复述句对[45].也有人使用句子聚类的方法抽取复述句,即首先计算各篇可比新闻的句子之间的相似度,然后在此基础上进行聚类.这里,句子相似度的计算可以基于词重叠率[44],也可以基于命名实体的重叠率[29]等.此类方法的优点在于互联网上可比新闻的数量众多,因此使用该方法可以构建相当规模的复述句库.这种方法的缺点在于其获取的复述句仅限于新闻体裁.另外,由于文本聚类、句子聚类等产生的错误级联,使得获取的复述句往往含有大量噪声,准确率较低.

此外,也有人将复述句获取定义为二元分类问题,即任意给定两个句子,由分类器判断二者是否为复述句[11,47−50].这种方法的关键在于分类特征的选择.具体地,有人使用两个句子的长度比、词重叠率、短语重叠率、命名实体重叠率、编辑距离以及同义词等特征计算这两个句子的相似度[47,48,50].也有人使用机器翻译自动评价中常用的 BLEU,NIST,WER(word error rate)和 PER(position-independent word error rate)等指标作为分类特征来判别两个句子是否为复述[49].表面上看,这种方法定义简单、明确,无须限定使用平行语料或可比语料,然而这种方法的实际效果却不甚理想.我们在之前的研究中也曾探索过此类方法[11].通过实验我们发现,其根本问题还在于特征的选择.由于语言的复述现象纷繁复杂,目前提出的种种特征很难准确并完全地描述和表征复述现象.通过对实际例子的观察,我们看到,两个看上去非常相似的句子可能因为某个细微的不同而使得意思截然相反;两个字面上差别迥异的句子却可能表达了相同的含义.所以,尝试使用更深层、更复杂的分类特征也许是这类方法的突破方向之所在.

除了前面介绍的几类主要方法以外,还有一些有意义的尝试值得我们关注.例如,Wu 将其提出的 ITG (inversion transduction grammar)模型用于复述句的识别[51];Brockett 等人开发了一个网络游戏,将编写复述设计为游戏情节,使得人们在休闲娱乐的同时提供复述句对[52].限于篇幅,本文对这些方法不作详细介绍.

2.2 复述短语与复述模板的获取

复述短语与复述模板的获取一直是复述研究方面最主要的内容.对此,前人提出了大量有价值的模型和方法,也尝试使用了多种数据和资源,其中既包括上面介绍的复述句库,也包括词典、双语资源、互联网等.鉴于复述短语与模板在获取方法与所需资源方面的相似性,本节对其进行统一介绍.

** http://www.nist.gov/speech/tests/mt/

© 中国科学院软件研究所 http://www.jos.org.cn

2.2.1 基于词典的复述获取

在本文中,我们将复述词视作复述短语的特例,而基于词典的方法则是获取复述词的典型方法.这里的词典特指类义词典(thesaurus),如WordNet*** 等.在类义词典中,词语按照各种词义关系加以组织,因此可以很方便地找到与一个词意思相近或相关的词.在基于词典的复述获取中,人们最容易想到的就是将词典中定义的同义词作为复述词抽取出来,这也是实际应用中最为普遍的一种做法[19,40].也有人突破这一约束,利用词典中定义的其他词义关系抽取复述,如WordNet中的attributes,pertainyms,seealsos等[34].根据这些词义关系抽取出来的词与原词的意思相关而不相同,但在某些具体的上下文语境中却可能相互替换.总的来说,上述方法可概括为将词典中定义的某种或某些种词义关系视作复述关系.然而究竟哪些词义关系可被视作复述却没有定论[1].除了WordNet等人工构造的词典之外,通过聚类的方法自动构造的同义词词典[53]也可用于复述词的抽取[10].

还有一种利用词典获取复述的方法,是将一个词在词典中的注释作为其复述抽取出来[54,55].如上所述,将专业术语等复述为其注释有助于人们的理解,因而这种复述有明确的应用目的.这里需要注意的问题是,一个词在词典中的注释可能是词或简单的短语(如“amnesia: memory loss”),也可能是比较复杂的短语,其中含有一些修饰限定内容(如“wrestle: combat to overcome an opposing tendency or force”).对于后者而言,在一个句子里直接将原词替换为注释通常会导致句法错误或句子不通顺.对此,研究者一般会通过总结词典注释的格式,切分出其中的主干,或者对词典注释进行某些转换和变形[54,55].我们在之前的研究中通过比较发现,Encarta在线词典**** 提供的注释比较简洁,更适合作为复述使用[10].

2.2.2 基于锚点(anchor)的复述获取

前面我们介绍了获取复述句对并构建复述句库的方法.利用构建的复述句库,我们可以进一步从中抽取复述短语和模板.为此,研究者提出了基于锚点的方法.该方法的基本思想是:在一对复述句中,通常存在部分相同的词和短语(如图 1(a)中标示下划线的部分所示)这些词和短语被称作锚点并加以对齐,而在对齐的锚点之间的内容(如图 1(a)中方框内的部分所示)即可被作为复述抽取出来.

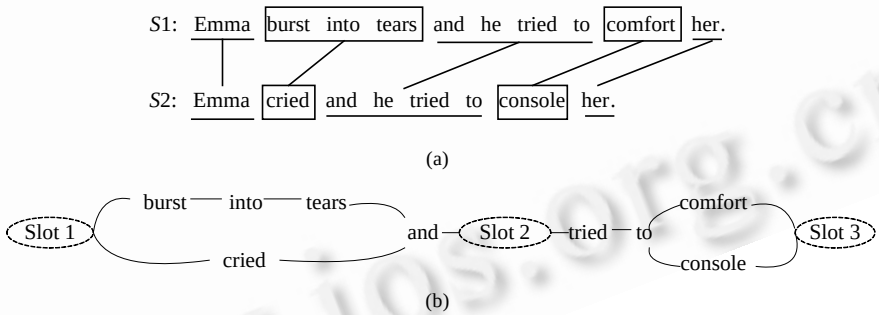

Fig.1 Example of paraphrase acquisition based on anchor approach

图 1 基于锚点的复述获取方法示例

Barzilay 和 McKeown 首先提出这种基于锚点的方法,并将其应用于复述短语的抽取[1].文献[1]对上面介绍的基本思想进行了改进,即不是简单地使用复述句对中的相同词作为锚点,而是基于 co-training 算法从复述句对中迭代地抽取复述及其周围的锚点特征(词性序列).

事实上,这种基于锚点的方法更多地用于复述模板的获取[29,42−44].简单来讲,就是将锚点对应的内容(或其中的部分内容)抽象成代表变量的槽,从而将复述短语保留下来.例如,对于图 1(a)中的例子,可以抽取出如图 1(b)所示的复述模板.由于篇幅限制,以上仅仅介绍了基于锚点的复述抽取方法的基本思想、大致过程和简单示例.在实际实现过程中,每种具体的方法都提出了不同的策略、算法和模型.例如,有的使用命名实体作为锚点[29],

*** http://wordnet.princeton.edu/

**** http://encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx

© 中国科学院软件研究所 http://www.jos.org.cn

有的采用多序列对齐的策略[44],还有的提出基于句法的锚点对齐[43]等.

2.2.3 基于枢轴(pivot)的复述获取

由于复述句库的规模和领域局限性较大,人们开始尝试使用更易获得且规模更大的双语语料来抽取复述短语和模板.为此,人们提出了基于枢轴的方法[4,56−58].该方法的基本假设是:一对复述短语虽然字面上表达不同,但其外文翻译却可能相同.例如,turn on the light 和 switch on the light 是一对英文复述短语,二者皆可翻译成中文短语“打开灯”.这样,人们就可以利用双语语料中提供的翻译信息识别复述短语.研究者将一组复述短语 $E=\{e_1,e_2,\ldots,e_n\}$ 对应的公共外文翻译 f 称为这组复述短语的枢轴[4].

最早的基于枢轴的方法比较简单,只是使用一部英-日翻译词典来获取英文复述词[56].即将对应相同日文翻译的英文词作为复述词抽取出来.后续研究者将基于枢轴的方法应用于大规模的双语平行语料库[58].他们首先使用机器翻译中的词对齐和短语抽取技术从双语平行语料库中抽取出短语翻译对,然后利用外文翻译作为枢轴抽取英文复述短语.设英文短语 e_1 和 e_2 对应的枢轴为 f,则该方法将 e_1 到 f 的翻译概率和 f 到 e_2 的翻译概率的乘积作为 e_1 到 e_2 的复述概率.受该研究的启发,我们将基于枢轴的方法用于复述模板的抽取[4].我们使用对数线性(log-linear)模型计算复述概率,抽取出像 consider X 和 take X into consideration 这样的复述模板.

Wu 和 Zhou 将类似的方法用于复述搭配的抽取[57].例如,〈solve,OBJ,problem〉和〈resolve,OBJ,problem〉就是一对互为复述的动宾搭配.该方法首先利用一个双语语料库为任意搭配获取中文翻译,然后将一个搭配的所有可能的翻译组成一个向量.若两个搭配的翻译向量的相似度很高,则这两个搭配将被作为同义搭配抽取出来.总之,上面介绍的种种基于枢轴的方法都说明了外文翻译信息在复述抽取中所起到的重要作用.

2.2.4 基于分布假设(distributional hypothesis)的复述获取

分布假设由语言学家 Zellig Harris 提出[59],其基本思想是那些倾向于出现在相似的上下文中的词意思相近.许多研究者基于分布假设计算词义相似度并实现同义词自动聚类[53].在复述研究中,有人对基本的分布假设进行扩展,即认为那些倾向于出现在相同的上下文中的短语或模板也应当意思相近[60].进而,研究者们基于这一扩展的分布假设抽取复述短语和模板[26,33,60−63].此类方法不需要基于复述句库或者双语平行语料库,而是需要一个大规模的单语语料库.有的研究者甚至直接将整个互联网作为其语料库,即使用网络挖掘的方法进行复述抽取[26,61,63].基于分布假设的复述获取方法还可细分为两类,一类是需要输入种子的方法(以下简记为方法 1),另一类则是不需要输入种子的方法(记为方法 2).

方法 1[26,61,63]常被用于对信息抽取(IE)模板的复述.对于信息抽取而言,通常需要首先定义类别体系,如生日、出生地、作者等,然后再针对每一类分别构建 IE 模板.当类别数十分巨大时,人们便需要一种自动生成 IE 模板的方法.对此,方法 1 首先针对每个类别手工构造少量实例,如生日类的“Mozart,1756”,“Mao Zedong,1893”等.这些实例即称为这个类别的种子.接下来,利用这些种子到语料库或互联网上搜索匹配的片段,抽取出每个种子之间的路径作为 IE 模板.例如,种子“Mozart,1756”可以匹配“born in 1756, Mozart”,“Mozart was born in 1756”,“Mozart (1756-”等片段,从中可以抽取出模板“born in Y, X”,“X was born in Y”,“X (Y-”等.这些模板即可视作复述模板.此类方法的关键在于对输入种子的选择.有人采用迭代的方式利用抽取的模板到语料库中匹配得到新的实例以扩充初始输入的种子[61].类似地,也有人直接输入给定类别的一个 IE 模板,利用这个初始的模板匹配得到若干实例,再利用这些实例作为种子获取更多的模板[63].

与方法 1 不同,方法 2[33,60,62]无须事先定义分类体系,也无须输入种子实例或模板.对于任意短语(或模板)e,方法 2 首先从语料库中统计出 e 的上下文,即 e 左、右各 n 个词,并用这些词构成 e 的上下文向量.根据扩展的分布假设,方法 2 通过计算两个短语(或模板)e_1 和 e_2 的上下文向量的相似度来度量二者的相似度.当相似度大于某一阈值时,即将其判定为复述.由于方法 2 在计算过程中需要存储大量短语或模板的上下文信息,因此其计算时间和空间复杂度均较高.

3 复述的生成

在以往的研究中,复述生成通常是指句子级复述生成.复述生成任务可定义为:给定一个句子 S,生成 S 的一

© 中国科学院软件研究所 http://www.jos.org.cn

个或一组复述句 $T=\{T_i|i\geq 1\}$.如前所述,复述生成可应用于机器翻译中的输入句改写、自动问答中的问句扩展,以及自然语言生成中的句子加工与润色等.因此,复述生成一直是复述研究中的一个热点问题.研究者们提出的复述生成方法可概括为以下几类:

3.1 基于规则的方法

复述生成的早期研究大多采用基于规则的方法[5,24,64].研究者针对特定应用(如自动问答中的问句改写[24])或特定问题(如状语位置变换、长句切分成短句等[5])制定复述生成规则.这里,复述生成规则的制定通常要基于一系列的语言分析与预处理,如词性标注、命名实体识别和句法分析等.相应地,制定的复述规则也包括词序列规则、词性规则以及句法规则等形式.如此生成的复述规则表达能力更强,实用效果也更好.此类方法最主要的优点在于针对性强,可以直接针对待解决的问题或特定的语言现象制定复述生成规则.其缺点在于规则的可扩展性差,人工编写规则的工作量大,成本较高.

3.2 基于词典的方法

基于词典的复述生成简单来讲就是将待复述句 S 中的某些词替换成其词典中定义的同义词,从而生成 S 的复述句 T 的过程.这一过程也被称为词替换(lexical substitution).近年来,国际上出现了针对词替换任务的评测.我们将在后文中对其进行详细介绍.这里,假设指定原句 S 中待替换的词为 w,则基于词典的复述生成通常包括两个主要步骤,即候选复述获取和复述确认.其中,候选复述获取是指从诸如 WordNet 等词典中抽取 w 的同义词作为其候选复述词.由于并非 w 的所有同义词在给定的上下文句 S 中均可替换 w,因此需要一个复述确认的步骤以过滤掉那些在 S 中不可替换 w 的候选复述词.复述确认的方法很多,比如,可以通过训练一个二元分类器来判别一个候选复述词在给定的上下文句中的可替换性[19].也可以通过统计替换后生成的短语或搭配在网络上的出现次数来判断该替换是否合理[40].这种基于词典的复述生成方法的优点在于简单易行,所需的同义词词典比较容易获得.其缺点在于生成的复述类型单一,仅能实现词替换而无法实现其他类型的复述变化.

3.3 基于模板的方法

基于模板进行复述生成的过程可概括为:设 D 是一个复述模板库,p_s 和 p_t 是 D 中的一对复述模板.若待复述句 S 匹配了 p_s,则将 p_s 各个槽中填充的内容填入 p_t 相应的槽中,从而生成复述句 T.此类方法的关键在于如何构建大规模的复述模板库.如上所述,已有很多研究工作着手解决面向复述生成的复述模板抽取问题[15,43,44,65].由于复述模板往往既包含词汇和短语的替换,又包含语序和句子结构的变化,因此基于模板的方法通常可以生成复杂而丰富的复述句.而此类方法存在的主要问题在于复述模板匹配的覆盖率较低,即能够匹配复述模板的句子较少.这也给复述模板抽取的研究提出了新的挑战,即如何控制复述模板的长度、复杂度和泛化程度,以使得在保证复述生成的准确率的同时提高复述模板匹配的覆盖率.

3.4 基于NLG的方法

对于传统的自然语言生成(NLG)系统而言,其输入是记录句法或语义信息的内部表示,而输出则是根据输入信息生成的自然语言句子.基于 NLG 的复述生成方法[66−68]主要包含两个步骤:步骤 1 是通过句法分析、语义分析等一系列处理得到待复述句 S 的内部表示 R;步骤 2 是基于 NLG 技术或直接利用现有的 NLG 系统生成 R 对应的自然语言句子 T.由于 S 和 T 都对应相同的中间表示 R,因此可保证 S 和 T 意思相同.同时,由于上述分析与生成这两个过程是相互独立的,因此可保证 T 与 S 在字面表达上的不同,即满足 T 是 S 的复述.基于 NLG 的方法模拟了人的复述行为,即先理解句子的意思,再用另外一种方式将其表达出来.然而,这种方法对深层语义知识的依赖性较大,通常需要建立一整套语义表达形式和框架.另外,NLG 系统的实现也十分复杂.

3.5 基于MT的方法

基于机器翻译(MT)的复述生成方法是将复述生成看作单语机器翻译,即源语和目标语为同一种语言的机器翻译.进而,研究者将比较成熟的统计机器翻译模型和系统应用到复述生成问题上来[46,69].与一般的统计机器翻译一样,基于 MT 的复述生成也需要一个大规模平行语料库作为训练语料,只不过该语料库是由复述句对而

© 中国科学院软件研究所 http://www.jos.org.cn

非翻译句对构成.这也正是该方法的瓶颈所在,因为复述句对的获取远比翻译句对的获取困难,在现实世界中很难找到大规模、高质量的复述句对资源.对此,我们在之前的研究中提出在 MT 框架下综合使用多种资源进行复述生成,以缓解前人的方法对大规模训练语料的依赖[10].

4 复述的评测

复述的评测一直是复述研究中的难点,目前国际上既没有针对复述的公共评测,也没有被广泛认可和接受的评测数据,因而研究者往往根据具体的研究问题,自行设计实验并标注实验数据.由于实验数据和评测指标不一致,研究者们提出的复述方法和模型很难具有可比性.本节将对前人提出的复述评测方法加以介绍.

4.1 针对复述获取和复述生成的评测

在机器翻译、自动文摘等相关研究中,自动评测方法都已相对成熟.然而对于复述而言,还是主要依赖人工评测.人工评测的弊端很多,其中最重要的是人工评测主观性强,不同评测者之间的差异性大.另外,评测标准往往难以制定.两个短语、模板或句子意思很难完全相同,那么到底相似到什么样的程度可以将其视为复述?已有文献中给出的标准也比较模糊,“意思大致相同”、“基本可互相替换”等在实际评测中很难把握[70].正因为如此,研究者在采用人工评测时通常都需要多个评测者,且之前要对评测者进行细致培训,之后还要考察评测者之间的一致性.这样做都是为了提高人工评测的客观性与公正性.与人工评测相比,复述的自动评测方法使用得较少,尚且处于辅助位置.以下是对各种具体评测方法的详细介绍.

4.1.1 对复述句获取的评测

对复述句获取的人工评测即是从获取的复述句对中随机选择若干对,然后由评测者标注每个句对是否确为复述.例如,Dolan 和 Brockett 请了 3 位评测者对其获取的复述句对进行标注,并通过互联网共享了该标注结果[48].除了人工评测之外,Dolan 等人还提出了一种基于词对齐的评测方法[45,47].该方法的基本假设是若一个复述句库 D 中的复述准确率较高,则在 D 之上的词对齐错误率应该较低;反之,若 D 中含有大量噪声,则对齐错误率会比较高.然而这种基于词对齐的方法也需要人工标注少量词对齐作为评测数据.我们认为,可以基于词对齐的方法进行复述句对的过滤,即首先对复述句库进行词对齐,然后过滤掉那些对齐结果差的句对.

4.1.2 对复述短语获取的评测

对复述短语获取的评测主要也是采用人工评测的方法[1,33,58,62,71,72].与句子相比,短语长度较短,歧义性较强,故而两个短语是否为复述往往与具体上下文相关.因此,有研究者在对复述短语进行评测时提供若干上下文句,评测者只需判断两个短语在给定的上下文句中是否为复述.这样一来,既降低了标注难度,又提高了标注准确性[1,58].Fujita 和 Sato 在对复述短语进行人工评测时细化了评测指标,评测者需要分别就语义是否等价、是否在某些上下文中可替换以及是否符合语法这 3 个指标对复述短语进行评测[71].这样做的好处在于使得评测结果更细致,也更直观.此外,Callison-Burch 等人提出了一种自动评测的方法[70].该方法使用一个指定的小规模语料库,首先人工从中抽取出全部可能的复述短语对作为标准集合,然后由各种待评测的复述抽取方法分别自动地从该语料库中抽取复述,最后通过对照自动抽取结果与标准集合,计算准确率和召回率.这种评测方法的弊端在于,各种复述抽取方法所使用的资源本不相同,因此限定所有方法使用给定的语料库抽取复述有失公平.另外,该评测方法只能计算各种复述获取方法的准确率下限和相对召回率,因此其价值有待考证.

4.1.3 对复述模板获取的评测

在对复述模板抽取的评测中,研究者同样广泛采用人工评测的方法[29,42,44,60].然而人们逐渐认识到,模板是对具体短语或句子的抽象,其中含有表示变量的槽,因此,仅仅对模板本身进行评测很困难,两个模板是否为复述应取决于槽中所填的具体内容,即需要将模板实例化.基于这一考虑,有人提出了基于实例的人工评测方法[73,74].具体来说,设 p_1 和 p_2 是待评测的两个模板,该评测方法首先找到可以匹配模板 p_1 的若干实例句,然后使用 p_2 对各实例句进行改写.若经过改写生成的新句子大部分是对应原句子的复述,则判定 p_2 是 p_1 的复述.这种评测方法有助于评测者作出准确判断,同时也可提高评测者之间的一致性.

© 中国科学院软件研究所 http://www.jos.org.cn

4.1.4 对复述生成的评测

对复述生成的评测类似于前面介绍的对复述句获取的评测,二者都是判断两个句子是否为复述.然而,对复述生成的评测更为困难,因为自动生成的复述句通常存在语病且不通顺,这会影响评测者的理解和判断.对复述生成的评测也是主要采取人工评测的方法[10,15,44,46].评测者既要考虑生成复述句的忠实度(即是否与原句意思相同),又要考虑生成复述句的自然度(即是否通顺、易懂)[15].目前还没有被研究者普遍接受的复述句人工评测规范和指导.相对而言,Finch 等人提出的评测方法比较全面和具体[69].他们分别提出了 3 种针对复述生成的评测方法.第 1 种是借鉴机器翻译的人工评测,由评测者为生成的复述句打分,根据复述句的质量,分数从 1 分到 5 分不等;第 2 种是借鉴机器翻译的自动评测方法 BLEU,首先构造候选答案,然后利用其进行自动打分;第 3 种是基于句法分析评测生成复述句的句法通顺度,其基本假设是句子的通顺度越差,句法分析产生的异常结果就越多.我们认为,Finch 等人提出的 3 种评测方法都很有借鉴价值.

4.2 面向应用的评测

面向应用的评测是指通过评测复述对于特定应用的贡献来间接地评测复述本身.例如,有人评测基于复述的 IE 模板扩展对于 QA 系统的贡献[26],有人评测通过使用复述技术解决数据稀疏能否使机器翻译的性能有所提高[17],还有人评测复述在关系抽取中的作用[32].这种评测方法将复述作为一个模块置于一个系统之中,以视其对整个系统性能的提高.然而,一个系统的性能往往受其所有模块的共同影响,其中复述模块是否能够起作用以及能够起多大作用都要受其他模块的制约.另外,在一种应用中被证明有价值,并不能保证其在其他应用中也有价值,也就是说,这种评测方法无法对复述整体上进行全面的评价[70].综合以上原因,研究者还是更倾向于对复述进行直接评测,而将这种面向应用的评测作为一种辅助手段.

5 蕴含与词替换

蕴含与词替换是与复述很相似的两个研究方向.下面,我们将对二者进行较为详细的介绍.我们期望能够从中借鉴一些有价值的思路和方法,从而对复述研究有所促进和启发.

5.1 蕴含(entailment)

蕴含是最近几年出现的研究方向,由 Dagan 和 Glickman 首先提出[75].蕴含被定义为一个文本(text)T 和一个假设(hypothesis)H 之间的二元关系,若 H 可以由 T 推理出来,则说 T 蕴含 H.我们通过一个例子对上面的定义加以解释.设 T 为“Jane is a French teacher”,H 为“Jane can speak French”,则 T 蕴含 H,因为 H 可从 T 中推理出来.但需要注意的是,蕴含中的推理是单方向的,即并不保证 T 也能从 H 中推理出来,上面的例子即可说明这一点.

蕴含与复述很相似,研究者一般将复述看作蕴含的特例,即给定文本T_1 和T_2,若T_1 蕴含T_2,同时T_2 也蕴含T_1,则T_1和T_2互为复述.换言之,蕴含研究的是语义的包含关系,而复述则研究的是语义的等价关系.与复述研究缺乏公共评测不同,国际上有针对蕴含识别的评测RTE(recognizing textual entailment).该评测从 2005 年开始举办[76],至今已举办 3 届(RTE-1~RTE-3)*****.由于RTE-3 中各参评系统代表了目前蕴含研究的最新进展和最高水平,因此,下面对RTE-3 评测中各参评系统的模型和方法作一总结.

最简单的一类识别蕴含的方法是将其定义为相似度计算问题,即计算 T 和 H 的相似度,当相似度的值超过特定阈值时,即判定蕴含关系成立[77,78].研究者一般会在多个维度上计算 T 和 H 的相似度,如词汇相似度、句法相似度等,然后再对各个相似度值进行融合.这种方法虽然简单,但实际评测结果却并不差,在全部参评系统中处于中游水平.

另外一类蕴含识别方法是基于机器学习的方法,这也是 RTE-3 评测中最普遍的一种方法.在这种方法中,蕴含识别被定义为二元分类问题,即分类器输入 T 和 H,输出 1(蕴含)或 0(非蕴含)[79−83].此方法也可以看作是对基

***** RTE-4 作为 TAC-2008(Text Analysis Conference)的一个子任务举行,截至本文成稿时评测尚未结束,因此本文不包含对 RTE-4 的相关介绍.

© 中国科学院软件研究所 http://www.jos.org.cn

于相似度方法的一种扩展,研究者们会将 T 和 H 的相似度作为分类特征使用.这种基于分类的方法的最大优点在于可以方便地融合多种特征,如词性特征、命名实体特征、句法特征、语义特征等.RTE-3 评测中很多系统都采用这一方法,但由于他们使用的特征和分类模型不同,因此其性能也相差迥异.

还有一种常见的蕴含识别方法是基于推理的方法[84,85].此类方法通常建立在对语言(T 和 H)的深入分析的基础上,人工编写细致而复杂的推理规则,并将这些推理规则用于从 T 到 H 的推理过程.与前面介绍的方法相比,这种基于推理的方法的一个显著缺点是不易实现,尤其是推理规则的编写既需要丰富的知识,又需要充足的人力.然而在 RTE-3 评测中成绩最好的两个系统均是基于此方法[84,85],而且其系统性能要远远优于其他参评系统.

由于复述与蕴含的相似性,二者在研究方法上的相通性也很强.因此,在今后的复述研究中应关注蕴含研究的发展动态,学习和借鉴其研究思路和方法.

5.2 词替换(lexical substitution)

前面我们曾介绍过,词替换是指将给定句子S中的词w替换成与其等价的词的过程.词替换也可看作是一种复述生成.在SemEval-2007 国际评测******中包含英文词替换的评测任务(任务 10)[86].该评测任务可分为两个步骤,即候选替换词获取以及替换词的选择.对于第 1 步,由于该评测对参评者使用的资源并无限定,因此各参评系统在获取候选替换词时实验了各类词汇资源,包括WordNet[87−90]、Encarta词典[87]牛津美国写作词典(Oxford American Writer Thesaurus)[88]等.其中,参评者主要使用各种词典中的同义词作为候选替换词[87,88,90],但也有人考察词典中定义的诸如SIMILAR-TO,ENTAILMENT等各种词汇关系[89].在替换词选择的步骤里,需要对所有候选替换词打分排序,并选出在给定上下文句中最适合的替换词.对此,大多数参评系统采用了网络挖掘的方法或使用大规模网络数据[91]来统计词替换后生成的句子片段的出现次数.其出现次数越多,说明该替换正确的可能性就越大[87−90].

词替换中的某些方法对于复述生成会有所启发,例如,如何选择获选复述、如何对生成的复述句打分排序等,因而词替换也是今后需要持续关注的一个研究方向.

6 结束语

本文在充分调研和深入分析的基础上对复述研究的进展进行了综述.其中重点介绍了复述研究中的几个关键问题,包括复述的应用、复述资源的获取、复述句生成、复述的评测等.虽然对复述问题的探讨由来已久,但在自然语言处理领域对其进行广泛研究却只有 10 年左右的时间.所以应该说复述研究中还存在许多值得深入探索的问题.在本文的最后,我们基于几年来的研究经验提出一些值得进一步挖掘的研究点,希望对本领域的其他研究者有所启发.

问题 1:虽然人们已经提出了多种方法用于获取复述句、复述短语和复述模板等资源.然而总的来看,获取的资源精确率较低,含有噪声太多,无法真正实用.因此,接下来的一个主要目标就是如何找到一种有效的方法,对各种复述资源进行过滤.尤其是希望能够找到一种对各类复述资源都适用的过滤方法.

问题 2:相对于复述资源获取,复述生成方面的研究还比较少,尤其是在生成模型上的研究还很不够.无论是基于 NLG 的模型还是基于 MT 的模型都是将复述生成看成其他问题的一个特例,并套用已有模型加以解决.然而,复述生成毕竟是一个独立的问题,有其区别于 NLG 和 MT 的特点,因此需要提出专门的复述生成模型.

问题 3:如前所述,复述的评测一直是困扰研究者的一个问题.因此希望广大研究者能够共同努力,推出一套大家都能认可和接受的评测体系和数据,这样一定会极大地促进本研究的发展.

问题 4:复述在众多相关研究中都有很大的应用潜力,但具体应用的例子却不多.这一方面是由于目前复述的精度欠缺,很多时候无法对相关研究起到促进作用;另一方面也是由于人们在这方面的尝试还不够.所以,在今后的研究中,应当积极探索复述在各种应用中的作用,这样才能更好地促进此项研究的深入开展.

****** http://nlp.cs.swarthmore.edu/semeval/

© 中国科学院软件研究所 http://www.jos.org.cn

致谢 在此,我们向对本研究工作提供帮助的老师和同学表示感谢.特别地,我们要感谢张志昌、贺瑞芳和蓝翔同学对本文初稿的审阅以及提出的宝贵意见.

References:

[1] Barzilay R, McKeown KR. Extracting paraphrases from a parallel corpus. In: Proc. of the ACL/EACL. Morristown: Association for Computational Linguistics, 2001. 50−57.

[2] Rinaldi F, Dowdall J, Kaljurand K, Hess M, MolláD. Exploiting paraphrases in a question answering system. In: Proc. of the IWP. Morristown: Association for Computational Linguistics, 2003. 25−32.

[3] Boonthum C. iSTART: Paraphrase recognition. In: Proc. of the ACL 2004 Workshop on Student Research. Morristown: Association for Computational Linguistics, 2004. 31−36.

[4] Zhao SQ, Wang HF, Liu T, Li S. Pivot approach for extracting paraphrase patterns from bilingual corpora. In: Proc. of the ACL 2008: HLT. Morristown: Association for Computational Linguistics, 2008. 780−788.

[5] Zong CQ, Zhang YJ, Yamamoto K, Sakamoto M, Shirai S. Approach to spoken Chinese paraphrasing based on feature extraction. In: Proc. of the NLPRS. 2001. 551−556.

[6] Zong CQ, Zhang YJ, Yamamoto K, Sakamoto M, Shirai S. Paraphrasing Chinese utterances in spoken language translation system. In: Proc. of the ICCC. 2001. 395−401 (in Chinese with English abstract).

[7] Li WG, Liu T, Zhang Y, Li S, He W. Automated generalization of phrasal paraphrases from the Web. In: Proc. of the IWP. 2005. 49−56.

[8] Liu T, Li WG, Zhang Y, Li S. 2006. Survey on paraphrasing technology. Journal of Chinese Information Processing, 2006,40(4): 25−33 (in Chinese with English abstract).

[9] Li WG. Research on Chinese paraphrase example and paraphrase template extraction [Ph.D. Thesis]. Harbin: Harbin Institute of Technology, 2008 (in Chinese with English abstract).

[10] Zhao SQ, Niu C, Zhou M, Liu T, Li S. Combining multiple resources to improve SMT-based paraphrasing model. In: Proc. of the ACL 2008: HLT. Morristown: Association for Computational Linguistics, 2008. 1021−1029.

[11] Zhao SQ, Zhou M, Liu T. Learning question paraphrases for QA from Encarta logs. In: Proc. of the IJCAI. Menlo Park: AAAI Press, 2007. 1796−1800.

[12] Zhao SQ, Liu T, Yuan XC, Li S, Zhang Y. Automatic acquisition of context-specific lexical paraphrases. In: Proc. of the IJCAI. Menlo Park: AAAI Press, 2007. 1789−1794.

[13] Mitamura T, Nyberg E. Automatic rewriting for controlled language translation. In: Proc. of the NLPRS. 2001. 1−12.

[14] Yamamoto K. Machine translation by interaction between paraphraser and transfer. In: Proc. of the COLING. Morristown: Association for Computational Linguistics, 2002. 1107−1113.

[15] Zhang YJ, Yamamoto K. Paraphrasing of Chinese utterances. In: Proc. of the COLING. Morristown: Association for Computational Linguistics, 2002. 1163−1169.

[16] Shimohata M, Sumita E, Y Matsumoto. Building a paraphrase corpus for speech translation. In: Proc. of the LREC. Paris: ELRA, 2004. 1407−1410.

[17] Callison-Burch C, Koehn P, Osborne M. Improved statistical machine translation using paraphrases. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2006. 17−24.

[18] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In: Proc. of the ACL. Morristown: Association for Computational Linguistics, 2002. 311−318.

[19] Kauchak D, Barzilay R. Paraphrasing for automatic evaluation. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2006. 455−462.

[20] Zhou L, Lin CY, Hovy E. Re-Evaluating machine translation results with paraphrase support. In: Proc. of the EMNLP. Morristown: Association for Computational Linguistics, 2006. 77−84.

[21] Lepage Y, Denoual E. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In: Proc. of the IWP. 2005. 57−64.

[22] Kanayama H. Paraphrasing rules for automatic evaluation of translation into Japanese. In: Proc. of the IWP. Morristown: Association for Computational Linguistics, 2003. 88−93.

[23] Madnani N, Ayan NF, Resnik P, Dorr BJ. Using paraphrases for parameter tuning in statistical machine translation. In: Proc. of the 2nd Workshop on Statistical Machine Translation. Morristown: Association for Computational Linguistics, 2007. 120−127.

[24] McKeown KR. Paraphrasing using given and new information in a question-answer system. In: Proc. of the ACL. Morristown:

© 中国科学院软件研究所 http://www.jos.org.cn

Association for Computational Linguistics, 1979. 67−72.
[25] Duboue PA, Chu-Carroll J. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2006. 33−36.
[26] Ravichandran D, Hovy E. Learning surface text patterns for a question answering system. In: Proc. of the ACL. Morristown: Association for Computational Linguistics, 2002. 41−47.
[27] Hermjakob U, Echihabi A, Marcu D. Natural language based reformulation resource and Web exploitation for question answering. In: Proc. of the TREC. 2002.
[28] Duclaye F, Yvon F. Learning paraphrases to improve a question-answering system. In: Proc. of the EACL Workshop on NLP for Question Answering. 2003.
[29] Shinyama Y, Sekine S, Sudo K. Automatic paraphrase acquisition from news articles. In: Proc. of the HLT. San Francisco: Morgan Kaufmann Publishers Inc., 2002. 40−46.
[30] Sekine S. Automatic paraphrase discovery based on context and keywords between NE pairs. In: Proc. of the IWP. 2005. 80−87.
[31] Sekine S. On-demand information extraction. In: Proc. of the ACL. Morristown: Association for Computational Linguistics, 2006. 731−738.
[32] Romano L, Kouylekov M, Szpektor I, Dagan I, Lavelli A. Investigating a generic paraphrase-based approach for relation extraction. In: Proc. of the EACL. Morristown: Association for Computational Linguistics, 2006. 409−416.
[33] Bhagat R, Ravichandran D. Large scale acquisition of paraphrases for learning surface patterns. In: Proc. of the ACL-08: HLT. Morristown: Association for Computational Linguistics, 2008. 674−682.
[34] Zukerman I, Raskutti B. Lexical query paraphrasing for document retrieval. In: Proc. of the COLING. Morristown: Association for Computational Linguistics, 2002. 1−7.
[35] McKeown KR, Barzilay R, Evans D, Hatzivassiloglou V, Klavans JL, Nenkova A, Sable C, Schiffman B, Sigelman S. Tracking and summarizing news on a daily basis with Columbia’s newsblaster. In: Proc. of the HLT. San Francisco: Morgan Kaufmann Publishers Inc., 2002. 280−285.
[36] Zhou L, Lin CY, Munteanu DS, Hovy E. ParaEval: Using paraphrases to evaluate summaries automatically. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2006. 447−454.
[37] Iordanskaja L, Kittredge R, Polguère A. Lexical selection and paraphrase in a meaning-text generation model. In: Paris CL, Swartout WR, Mann WC, eds. Natural Language Generation in Artificial Intelligence and Computational Linguistics. 1991. 293−312.
[38] Knight K, Chander I. Automated postediting of documents. In: Proc. of the AAAI. Menlo Park: AAAI Press, 1994. 779−784.
[39] Carroll J, Minnen G, Pearce D, Canning Y, Devlin S, Tait J. Simplifying text for language-impaired readers. In: Proc. of the EACL. Morristown: Association for Computational Linguistics, 1999. 269−270.
[40] Bolshakov IA, Gelbukh A. Synonymous paraphrasing using WordNet and Internet. In: Proc. of the NLDB. Berlin, Heidelberg: Springer-Verlag, 2004. 312−323.
[41] Uzuner Ö, Katz B, Nahnsen T. Using syntactic information to identify plagiarism. In: Proc. of the 2nd Workshop on Building Educational Applications Using NLP. Morristown: Association for Computational Linguistics, 2005. 37−44.
[42] Ibrahim A, Katz B, Lin J. Extracting structural paraphrases from aligned monolingual corpora. In: Proc. of the IWP. Morristown: Association for Computational Linguistics, 2003. 57−64.
[43] Pang B, Knight K, Marcu D. Syntax-based alignment of multiple translations: Extracting Paraphrases and Generating New Sentences. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2003. 102−109.
[44] Barzilay R, Lee L. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2003. 16−23.
[45] Dolan B, Quirk C, Brockett C. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proc. of the COLING. Morristown: Association for Computational Linguistics, 2004. 350−356.
[46] Quirk C, Brockett C, Dolan W. Monolingual machine translation for paraphrase generation. In: Proc. of the EMNLP. Morristown: Association for Computational Linguistics, 2004. 142−149.
[47] Brockett C, Dolan WB. Support vector machines for paraphrase identification and corpus construction. In: Proc. of the IWP. 2005. 1−8.
[48] Dolan WB, Brockett C. Automatically constructing a corpus of sentential paraphrases. In: Proc. of the IWP. 2005. 9−16.
[49] Finch A, Hwang YS, Sumita E. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In: Proc. of the IWP. Morristown: Association for Computational Linguistics, 2005. 17−24.
[50] Hatzivassiloglou V, Klavans JL, Eskin E. Detecting text similarity over short passages: Exploring linguistic feature combinations

© 中国科学院软件研究所 http://www.jos.org.cn

via machine learning. In: Proc. of the EMNLP. Morristown: Association for Computational Linguistics, 1999. 203−212.

[51] Wu DK. Recognizing paraphrases and textual entailment using inversion transduction grammars. In: Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. Morristown: Association for Computational Linguistics, 2005. 25−30.

[52] Brockett C, Dolan WB. Echo chamber: A game for eliciting a colloquial paraphrase corpus. In: Proc. of the KCVC. Menlo Park: AAAI Press, 2005. 8−15.

[53] Lin DK. Automatic retrieval and clustering of similar words. In: Proc. of the COLING/ACL. Morristown: Association for Computational Linguistics, 1998. 768−774.

[54] Kaji N, Kawahara D, Kurohash S, Sato S. Verb paraphrase based on case frame alignment. In: Proc. of the ACL. Morristown: Association for Computational Linguistics, 2002. 215−222.

[55] Higashinaka R, Nagao K. Interactive paraphrasing based on linguistic annotation. In: Proc. of the COLING. Morristown: Association for Computational Linguistics, 2002. 1218−1222.

[56] Takao K, Imamura K, Kashioka H. Comparing and extracting paraphrasing words with 2-way bilingual dictionaries. In: Proc. of the LREC. Paris: ELRA, 2002. 1016−1022.

[57] Wu H, Zhou M. Synonymous collocation extraction using translation information. In: Proc. of the ACL. Morristown: Association for Computational Linguistics, 2003. 120−127.

[58] Bannard C, Callison-Burch C. Paraphrasing with bilingual parallel corpora. In: Proc. of the ACL. Morristown: Association for Computational Linguistics, 2005. 597−604.

[59] Harris ZS. Distributional structure. In: Martinet A, Weinreich U, eds. Linguistics Today. New York: Linguistic Circle of New York, 1954. 26−42.

[60] Lin DK, Pantel P. Discovery of inference rules for question answering. Natural Language Engineering, 2001,7(4):343−360.

[61] Brin S. Extracting patterns and relations from the World Wide Web. In: Proc. of the WebDB'98. Berlin, Heidelberg: Springer-Verlag, 1998. 172−183.

[62] Pasca M, Dienes P. Aligning needles in a haystack: Paraphrase acquisition across the Web. In: Proc. of the IJCNLP. Berlin, Heidelberg: Springer-Verlag, 2005. 119−130.

[63] Szpektor I, Tanev H, Dagan I, Coppola B. Scaling Web-based acquisition of entailment relations. In: Proc. of the EMNLP. Morristown: Association for Computational Linguistics, 2004. 41−48.

[64] Takahashi T, Iwakura T, Iida R, Fujita A, Inui K. KURA: A transfer-based lexico-structural paraphrasing engine. In: Proc. of the NLPRS. 2001. 37−46.

[65] Fujita A, Inui K. A class-oriented approach to building a paraphrase corpus. In: Proc. of the IWP. 2005. 25−32.

[66] Power R, Scott D. Automatic generation of large-scale paraphrases. In: Proc. of the IWP. 2005. 73−79.

[67] Fujita A, Inui K, Matsumoto Y. Exploiting lexical conceptual structure for paraphrase generation. In: Proc. of the IJCNLP. Berlin, Heidelberg: Springer-Verlag, 2005. 908−919.

[68] Kozlowski R, McCoy KF, Vijay-Shanker K. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In: Proc. of the IWP. Morristown: Association for Computational Linguistics, 2003. 1−8.

[69] Finch A, Watanabe T, Akiba Y, Sumita E. Paraphrasing as machine translation. Journal of Natural Language Processing, 2004, 11(5):87−111.

[70] Callison-Burch C, Cohn T, Lapata M. ParaMetric: An automatic evaluation metric for paraphrasing. In: Proc. of the COLING. Morristown: Association for Computational Linguistics, 2008. 97−104.

[71] Fujita A, Sato S. A probabilistic model for measuring grammaticality and similarity of automatically generated paraphrases of predicate phrases. In: Proc. of the COLING. Morristown: Association for Computational Linguistics, 2008. 225−232.

[72] Glickman O, Dagan I. Identifying lexical paraphrases from a single corpus: A case study for verbs. In: Proc. of the RANLP. 2003.

[73] Pantel P, Bhagat R, Coppola B, Chklovski T, Hovy E. ISP: Learning inferential selectional preferences. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics, 2007. 564−571.

[74] Szpektor I, Shnarch E, Dagan I. Instance-Based evaluation of entailment rule acquisition. In: Proc. of the ACL. Morristown: Association for Computational Linguistics, 2007. 456−463.

[75] Dagan I, Glickman U. Probabilistic textual entailment: Generic applied modeling of language variability. In: Proc. of the PASCAL. 2004.

[76] Dagan I, Glickman O, Magnini B. The PASCAL recognising textual entailment challenge. In: Proc. of the MLCW 2005. Berlin, Heidelberg: Springer-Verlag, 2006. 177−190.

[77] Ferrandez O, Micol D, Munoz R, Palomar M. A perspective-based approach for solving textual entailment recognition. In: Proc. of

© 中国科学院软件研究所 http://www.jos.org.cn

the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 66–71.
[78] Wang R, Neumann G. Recognizing textual entailment using sentence similarity based on dependency tree skeletons. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 36–41.
[79] Malakasiotis P, Androutsopoulos I. Learning textual entailment using SVMs and string similarity measures. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 42–47.
[80] Ferres D, Rodriguez H. Machine learning with semantic-based distances between sentences for textual entailment. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 60–65.
[81] Montejo-Ráez A, Perea JM, Martínez-Santiago F, García-Cumbreras MÁ, Martín-Valdivia M, Ureña-López A. Combining lexical-syntactic information with machine learning for recognizing textual entailment. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 78–82.
[82] Adams R, Nicolae G, Nicolae C, Harabagiu S. Textual entailment through extended lexical overlap and lexico-semantic matching. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 119–124.
[83] Li BL, Irwin J, Garcia EV, Ram A. Machine learning based semantic inference: Experiments and Observations at RTE-3. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 159–164.
[84] Tatu M, Moldovan D. COGEX at RTE3. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 22–27.
[85] Hickl A, Bensley J. A discourse commitment-based framework for recognizing textual entailment. In: Proc. of the Workshop on Textual Entailment and Paraphrasing. Morristown: Association for Computational Linguistics, 2007. 171–176.
[86] McCarthy D, Navigli R. SemEval-2007 Task 10: English lexical substitution task. In: Proc. of the SemEval 2007. Morristown: Association for Computational Linguistics, 2007. 48–53.
[87] Hassan S, Csomai A, Banea C, Sinha R, Mihalcea R. UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In: Proc. of the SemEval 2007. Morristown: Association for Computational Linguistics, 2007. 410–413.
[88] Giuliano C, Gliozzo A, Strapparava C. FBK-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In: Proc. of the SemEval 2007. Morristown: Association for Computational Linguistics, 2007. 145–148.
[89] Martinez D, Kim SN, Baldwin T. MELB-MKB: Lexical substitution system based on relatives in context. In: Proc. of the SemEval 2007. Morristown: Association for Computational Linguistics, 2007. 237–240.
[90] Zhao SQ, ZhaoL, Zhang Y, Liu T, Li S. HIT: Web based scoring method for English lexical substitution. In: Proc. of the SemEval 2007. Morristown: Association for Computational Linguistics, 2007. 173–176.
[91] Brants T, Franz A. Web 1T 5-gram Version 1. Technical Report, Philadelphia: Linguistic Data Consortium, 2006.

附中文参考文献:

[6] 宗成庆,张玉洁,山本和英,坂本仁,白井谕.口语自动翻译系统中的汉语语句改写.见:中文计算国际会议(ICCC).2001.395–401.
[8] 刘挺,李维刚,张宇,李生.复述技术研究综述.中文信息学报,2006,40(4):25–33.
[9] 李维刚.中文复述实例与复述模板抽取技术研究[博士学位论文].哈尔滨:哈尔滨工业大学,2008.

赵世奇(1981－),男,辽宁抚顺人,博士生,CCF 学生会员,主要研究领域为复述,自然语言处理.

李生(1943－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为自然语言处理,信息检索,机器翻译.

刘挺(1972－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息检索.

© 中国科学院软件研究所 http://www.jos.org.cn