·人工智能及识别技术· 文章编号: 1000-3428(2018)04-0231-05 文献标志码: A 中图分类号: TP18

基于语序变换的藏文复述句生成方法

柔 特[1a,1b],才让加[1a,1b],孙茂松[2]

(1.青海师范大学 a.计算机学院,西宁 810008; b.藏文信息处理教育部重点实验室,西宁 810008;
2.清华大学 计算机科学与技术系 智能技术与系统国家重点实验室,北京 100084)

摘 要: 机器理解藏文语句存在灵活性差和复杂性高的问题。为此,针对藏文相同语义句子的不同表达方式,设计复述句自动生成方法。通过对藏文句型结构、句子内部组块进行分析,利用全排列递归算法生成复述句。实验结果显示,与其他语言复述生成方法不同,该方法根据藏文句子中组块数量的不同,通过一个句子可以生成一个或多个,甚至上千个句义相同的复述句并且准确率达到93.4%,可应用于藏汉机器翻译、机器翻译评测和藏文问答系统等领域。

关键词: 复述生成; 藏文; 语序变换; 句型结构; 组块分析

中文引用格式: 柔 特,才让加,孙茂松.基于语序变换的藏文复述句生成方法[J].计算机工程,2018,44(4):231-235.
英文引用格式: ROU Te,CAI Rangjia,SUN Maosong. Tibetan Paraphrase Sentence Generation Method Based on Word Order Transformation[J]. Computer Engineering,2018,44(4):231-235.

Tibetan Paraphrase Sentence Generation Method Based on Word Order Transformation

ROU Te[1a,1b],CAI Rangjia[1a,1b],SUN Maosong[2]

(1a. Computer College; 1b. Key Laboratory of Tibetan Information Processing of Ministry of Education, Qinghai Normal University, Xining 810008, China; 2. State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

【Abstract】 Aiming at the problem of the flexibility and complexity of machines to understand natural language and Tibetan sentences, in view of the different expressions of the same semantic sentences in Tibetan language, this paper proposes a Tibetan paraphrases sentence generation method. Through the parsing of the sentence structure of Tibetan and the internal chunks of sentences, it uses permutation recursive algorithm to generate paraphrases sentence. Experimental results show that different from other languages and Tibetan chunks, the number of chunks in a sentence can generate one or more or even thousands of complex sentences with the same semantic meanings by the proposed method, and the accuracy of automatic generation of Tibetan paraphrases sentences reaches 93.4%. It can be applied to Tibetan-Chinese machine translation, machine translation evaluation, Tibetan QA system and other research fields.

【Key words】 paraphrase generation; Tibetan; word order transformation; sentence structure; chunk parsing
DOI: 10.3969/j.issn.1000-3428.2018.04.037

0 概述

自然语言理解是语言信息处理和人工智能领域的核心研究课题之一。判别计算机是否能理解自然语言必须具备4个标准,分别是问答系统、自动文摘、复述技术和机器翻译。计算机只要达到以上4条中的任何一种要求,就可以说它理解了自然语言。其中,复述技术是一个句子或短语转换成相同语义的句子或短语的技术。文献[1]将复述看作传达相同信息的可替换形式,文献[2]则认为复述反映一个语言的多变性,表示对应到相同意义的等价表达方式。文献[3-4]的定义则是概念上的近似等价。复述的简单解释就是对相同语义的不同表达[5]。

复述技术可应用在自动文摘、文本生成、信息抽取、自动问答、信息检索、机器翻译、情感分析[6-7]等领域。近年来,复述作为自然语言理解的一个重要

基金项目: 国家自然科学基金(61662061); 国家社会科学基金(14BYY132,16YY167); 教育部长江学者和创新团队发展计划项目(IRT1068); 青海省重点实验室项目(2015-Z-Y03,2017-GX-146)。
作者简介: 柔 特(1975—),男,副教授、博士研究生,主研方向为藏文信息处理; 才让加、孙茂松,教授、博士生导师。
收稿日期: 2017-01-09 修回日期: 2017-03-19 E-mail: crpengcuo13@yahoo.com

的研究方向，已成为学术界关注的热点。微软研究院、谷歌研究院、南加州大学、康奈尔大学等研究机构对英语的复述进行深入研究，所提出的方法与具体语言无关，可扩展到其他语种。东京大学、京都大学、ATR 研究院等机构对日文复述展开了研究，主要涉及日文特有的语言现象和特殊处理，该方法的语言相关性较强。国内对汉语复述的研究集中在哈尔滨工业大学，主要内容为词汇级复述[8]、短语级复述[9]、针对机器翻译的复述等[10]，研究方法与以上 2 种语言相似。此外，其对汉语复述资源获取、复述生成以及复述应用等方面也做了很多尝试。

目前关于国内少数民族语言方面的复述研究较少，特别地，研究者对藏文复述研究领域更少涉及。因此，本文在参考英语、日语和汉语复述研究成果和藏文词法分析的基础上，提出一种利用全排列递归算法生成藏文复述句的方法。

1 相关研究

对藏文句子级复述的研究方法可分为 2 个部分：1）获取资源，通过现有的资源构建复述句库；2）复述生成，通过输入的信息自动生成语义相同的复述句。

在复述资源获取方面，比较通用方法是当一个著作有多个版本的译文时，将不同译文版的句子对齐作为复述句[11]，此方法准确率高但资源有限。有些研究人员利用同一个主题不同媒体报道的新闻来做复述，也叫可比性语料[12-14]，虽然其核心主题相同，但文本长度和内容表达都各有千秋，比较适合于段落层面的复述。另外，也有研究者应用机器翻译的经验来获得复述，以及机器翻译评测的参考答案构建复述句库[15]。上述这些方法简单易行，但获得的复述句数量和领域有限。

对复述生成方法进行归类可得：早期多数研究者采用基于规则的方法[16-18]，此方法在特定条件下效果很好，但存在可扩展性差、工作量大、覆盖面窄等缺点。有研究者利用基于词典的方法将原句中的某些词替换成词典中的同义词或释义来生成复述句，此方法比较通用，生成效果较好。也有研究者利用机器翻译生成复述句，也就是将复述生成看作单语机器翻译。此外，也有研究者将统计机器翻译模型应用于复述生成[19-20]以及将平行语料库应用于多个翻译系统获得一对多译文结果[15]。

复述生成结果按形式划分可以分为 2 种：异形同义和同形同义。异形同义是指原文句子和复述句子之间字形不同但语义相同，主要通过同义词、释义、短语、句子结构的变化，利用从句、拆分与合并等技术手段生成复述句。同形同义是指原文句子和复述句子的组成成分完全相同，通过语序变换的方法生成复述句。

2 基于语序变换的复述生成

藏文语序变换的复述生成是在不改变原文句义的前提下，变换句子成分的位置但不改变句子的组成部件，在变换句子成分的位置时，组成谓语部件的位置必须在句末，其他句子成分的位置都比较灵活，可以出现在句首、句中。换而言之，改变这些成分的位置通常不会影响原文的句义。

句子是动态的话语运用单位，而句型是静态的语言模型。从句子结构看，藏文也可以分为简单句型、并列句型、复合句型。本文主要针对基于语序变换的藏文简单句型复述方法进行研究。

2.1 藏文句型的特点

藏文是藏族使用的文字，已有 1 300 多年的历史，据藏文史籍记载，藏文在历史上曾进行过 3 次较大规模的厘定规范。在吞弥·桑布扎时期语言文法著作有 8 种，如今只传世《三十颂》和《性入法》2 种。

藏文文法《三十颂》指出“字成词，词成句，句大意”。在藏文句子的语序中，谓语一般位于宾语之后，处在单句末尾[21]。当句末有动词构成的谓语时，动词后面没有“ཏུ་”等词缀表示一个完成的句子[22]。藏文句义的表达主要由虚词来限定，主语后面有虚词，句中实词之间的位置变换对整句表达影响不大，但谓语部分在句末[23-24]。例如：“ཁས་སང་བག་ཅིས་ཀྱིས་ཐང་ཀ་ཇིས།”（昨天扎西画了唐卡），该句的谓词部分“ཇིས”位于句末，所以，除该句子的谓词部分外，其他组成成分或语序可以变化。谓词成分在原位置不变的情况下，可生成不同语序变化且句义相同的句子，例如“ཁས་སང་ཐང་ཀ་བག་ཅིས་ཀྱིས་ཇིས།བག་ཅིས་ཀྱིས་ཐང་ཀ་ཡིན།”。

2.2 复述生成算法

在藏文复述生成过程中句子组块识别与复述生成方法是 2 个重要的环节。藏文组块分析是复述生成的预处理，通过简化句子结构为生成复述句提供基础。复述生成是通过句子构成部件或组块的语序变换生成一个或多个与原句同义的复述句的过程。

2.2.1 藏文句子组块生成模板原则

根据文献[25]对组块的定义，组块是一种语法结构，是符合一定语法功能的非递归短语。每个组块都有一个中心词，组块内的所有成分都围绕该中心词展开，任何一种类型的组块内部不包含其他类型的组块。

句子是语言运用的基本单位，由词、词组（短语）

构成，表达一个完整的意思。本文针对藏文句法分布和语法功能的特点，利用语言分析中常用的直接成分分析法，对藏文句子中的构成部件进行标识，从而标识词在不同句子成分中的语法功能信息，得到藏文组块，即得到藏文组块生成模板。例如："[དེ་རིང་] [བཀྲ་ཤིས་ཀྱིས་] [དགེ་ཚན་ལ་] [ལས་ཀྱི་] [དུས་ཐོག་ཏུ་] [སྐྱོ།]"（今天扎西把作业按时交给了老师），对其进行组块识别后可得"[དེ་རིང་] [བཀྲ་ཤིས་ཀྱིས་] [དགེ་ཚན་ལ་] [ལས་ཀྱི་] [དུས་ཐོག་ཏུ་] [སྐྱོ།]"。藏文组块生成模板实例如表 1 所示。

表 1　藏文组块生成模板实例

实例	块组合模板	实例	块组合模板
རྒྱ་ཡིག་དཔེ་ཆ	N+N	འགྲོ་དགོས	V+VC/དགོས
དཔེ་ཆ་བློག	N+V	ཁོང་གི	RR+འབྲེལ་སྒྲ/ཕྱིད་སྒྲ
མེ་ཏོག་དམར་པོ	N+A	གཞེར་ཕྱི་ནུ་རྒྱན	N+འབྲེལ་སྒྲ+N
ཡུག་གཞིག	N+M	བཀྲ་ཤིས་ཀྱིས་སྨོལ་མར	RR+ཕྱིད་སྒྲ+RR+ལ་དོན
མར་རྒྱ་མ་གང	N+Q+M	མཚོས་ཁྱིང་རིང	A+ཁྱིང+A
སློབ་ཁང་ནང་དུ	N+F	མགོ་ནག་མེ་ལ	N+N+ལ་དོན
དེ་རིང་ཐུ་དྲུ	T+T	སྐྱོལ་མས་ཁྱོ་རེས	RR+ཕྱིད་སྒྲ+N+ཕྱིད་སྒྲ
རྒྱ་མཚོ་ཟུ	N+ཚོག་ཐང	དོན་འགྲུབ་ཀྱི་ལག་ཏུ	RR+འབྲེལ་སྒྲ+N+ལ་དོན
སློབ་སྒྲོང་ཕྱིད	NV+ཕྱིད་དགོས		

藏文组块识别作为生成复述句的一种预处理手段，主要功能是在不需要深层语言知识的前提下，识别句子中特定的组块，如基本名词短语、时间词短语、代词短语、动词短语等，组块分析的目的是找出词、短语等的相互关系以及各自在句子中的作用，这种层次结构可以是从属关系、直接成分关系，也可以是语法功能关系。

2.2.2　复述句生成方法

输入一句藏文，输出一个或多个句义相同的藏文复述句，生成复述的前提条件是句子中必须有谓语部分，且谓语部分在句末。以"བཀྲ་ཤིས་ཀྱིས་སློག་ཚད་ནང་དུ་སོད་དུ་སོད་ཡིག་གཏགས།"（扎西在电脑上录入藏文）为例，生成藏语复述句的具体步骤如下：

步骤 1　对藏文原句分词　"བཀྲ་ཤིས་/ཀྱིས་/སློག་ཚད་/ནང་དུ་/སོད་ཡིག་/གཏགས་/།"。

步骤 2　按藏文组块生成原则生成藏文组块："[བཀྲ་ཤིས་ཀྱིས་] [སློག་ཚད་ནང་དུ་] [སོད་ཡིག་] [གཏགས།]"。

步骤 3　识别谓语组块。藏文组块生成结果中识别谓语部分　"[གཏགས།]"。

步骤 4　对非谓语组块进行全排列：

$$S=\{[\text{བཀྲ་ཤིས་ཀྱིས་}],[\text{སློག་ཚད་ནང་དུ་}],[\text{སོད་ཡིག་}]\Longrightarrow$$

$$\{w_1,w_2,w_3\}\Rightarrow\{w_1,w_3,w_2\}\Rightarrow\{w_2,w_1,w_3\}\Rightarrow$$

$$\{w_2,w_1,w_3\}\Rightarrow\{w_3,w_1,w_2\}\Rightarrow\{w_3,w_2,w_1\}$$

S 中有 3 个组块：w_1 {བཀྲ་ཤིས་ཀྱིས་}，w_2 {སློག་ཚད་ནང་དུ་}，w_3 {སོད་ཡིག་}，共 $3!=6$ 种排列，对集合 S 的全排列过程如图 1 所示。其中，S_O 表示原句子，S_P 表示 S_O 的复述句。藏文语序变换复述通常需要满足以下条件：1）S_O、S_P 为同一种语言，且句子构成成分完全相同；2）S_O、S_P 是结构上稳定的简单句；3）S_O、S_P 所表达的含义相同。

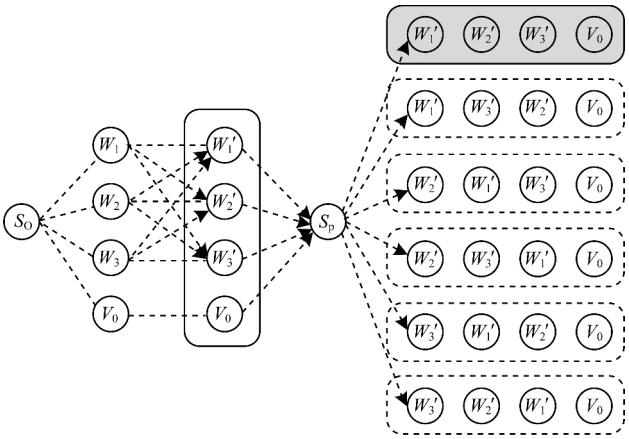

图 1　复述句生成示意图

步骤 5　通过尾部添加谓语组块生成藏文复述句。

图 1 中每条虚线表示该句子与句子构成部件之间的关系。例如：$\{W_1,W_2,W_3,W_4,V_O\}$ 是原文中组块元素的个数，$\{W_1',W_2',W_3',V_O\}$ 是每一次语序变换后生成的元素位置。在基于语序变换的复述生成过程中，左边的 $\{W_1,W_2,W_3,W_4,V_O\}$ 可以看作输入元素的个数。最右边的图是复述生成结果，包含原句 $A_n^m=n(n-1)(n-2)\cdots(n-m+1)$，其中，$A$ 为排列公式，n 为句子中构成组块的总数，m 为可循环组合的元素数，但不包含谓词部分的组块。以表 2 中的实例为例，其复述生成结果为 $A_3^3=3(3-1)\times(3-2)\times(3-3+1)=6$。藏文复述生成实例如表 2 所示。

表 2　藏文复述生成实例

实例	集合	备注
བཀྲ་ཤིས་ཀྱིས་སློག་ཚད་ནང་དུ་སོད་ཡིག་གཏགས།	S_O	藏文原句
བཀྲ་ཤིས་ཀྱིས་སོད་ཡིག་སློག་ཚད་ནང་དུ་གཏགས།	S_{P1}	
སློག་ཚད་ནང་དུ་བཀྲ་ཤིས་ཀྱིས་སོད་ཡིག་གཏགས།	S_{P2}	
སློག་ཚད་ནང་དུ་སོད་ཡིག་བཀྲ་ཤིས་ཀྱིས་གཏགས།	S_{P3}	藏文复述句
སོད་ཡིག་སློག་ཚད་ནང་དུ་བཀྲ་ཤིས་ཀྱིས་གཏགས།	S_{P4}	
སོད་ཡིག་བཀྲ་ཤིས་ཀྱིས་སློག་ཚད་ནང་དུ་གཏགས།	S_{P5}	

3　实验与结果分析

本文实验从小学藏语文教材语料中抽取 12 355 条句子，过滤掉词长 $d>15$（句子过长影响复述句子生成质量）的句子后得到 6 027 句，从中随机抽取 500 句作为实验用语料，对自动生成基于语序变换的藏文复述准确性进行考查。

在实验过程中，评测方法是对人工复述句和系统自动生成藏文复述句进行比较。首先对 500 句实验用语料按照本文给出的组块生成原则和复述句生成方法建立人工复述句库，给出原句的所有复述形式。计算机自动生成复述过程中先对 500 个藏文原句进行分词（分词工具使用了青海师范大学开发的班智达分词系统），将分词好的文件读入程序中，使计算机自动生成藏文复述句。

3.1 评测标准

设 A 为系统正确生成的复述数目，B 为系统自动生成复述的总数目，C 为人工句子复述数目，则实验的评测指标，即准确率 P、召回率 R 和综合评价 F 值计算公式如下：

$$P=\frac{A}{A+B}\cdot 100\%$$

$$R=\frac{A}{A+C}\cdot 100\%$$

$$F=\frac{2\cdot P\cdot R}{P+R}\cdot 100\%$$

人工与系统自动生成复述句实验结果对比如表 3所示。

表 3 人工与自动生成复述句实验结果对比

原句数量	人工生成数量	自动生成数量	正确生成数量	P/%	R/%	F/%
500	1 486	1 591	1 486	93.4	100	96.6

3.2 结果分析

本文以简单句型为例对基于语序变换的藏文复述句自动生成进行分析。从实验结果可以看出，对于500 个原句子系统总共自动生成了 1 591 个复述句，其中包含了所有人工复述句。最终计算机自动生成藏文复述句的准确率为 93.4%，召回率为 100%，F 值为 96.6%。同时可以实验结果中发现，原句的句型结构、组块数量，组块生成模板对复述生成结果有显著的影响。

3.2.1 句型结构对复述句生成结果的影响

在简单句型中主语和谓语是句子的主干，是句子的核心。简单句型的基本形式是由一个主语加一个谓语构成，可归纳为 5 个基本句型：主语 + 表语 + 系动词（S + P + V），主语 + 谓语（S + V），主语 + 宾语 + 谓语（S + DO + V），主语 + 双宾语 + 谓语（S + IO + DO + V），主语 + 宾语 + 宾补 + 谓语（S + DO + OC + C）。进一步对实验数据进行分析后发现，句型结构与复述句生成数量存在如图 2 所示关系。

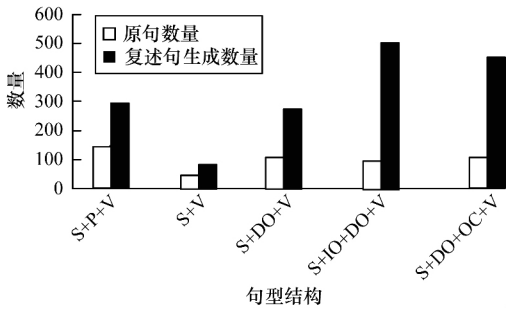

图 2 句型结构与复述句生成数量的关系

由图 2 可以看出，句型结构不同生成的复述句数量也存在较大差距。其中，S + IO + DO + V 句型结构是语序变换的藏文复述句生成过程中最理想一种句型结构，统计结果发现该句型结构的复述句生成数量超过其他结构。从实验数据得出，复述句生成最少的句型是 S + V，如“ཁ་བ་བབས།”（下雪了）和“ཉི་ཟེར་སྐྱེགས་བཞིར་འཕྲོས།”（太阳在照耀着）。此外，S + P + V 的复述生成数也较低，如“ཚེ་རིང་ནི་ཐུ་བོར་བ་ཞིག་ཡིན།”（才让是一位画家）和“འདི་ནི་སློབ་དེབ་ཡིན།”（这是教科书）。上述列举中前 2 个不能生成复述句，后两个只能生成 1∶1 的复述句，例如“སྐྱེགས་བཞིར་ཉི་ཟེར་འཕྲོས།”（太阳在照耀着）和“སློབ་དེབ་ནི་འདི་ཡིན།”（这是教科书）。

3.2.2 组块数量对复述句生成结果的影响

实验结果显示，复述句生成数量不在于句子长度而在于原句中组块数量的多与少。换言之，句中组块越多生成的复述句就越多，随着组块数的递增复述生成以阶乘式递增。

句型结构不同生成的组块数量不同。通过实验数据分析可知，上述 5 种基本句型中相对生成较多结构是 S + IO + DO + V 和 S + DO + OC + V。在实验语料库中每个句型结构与组块数量的平均分布如图 3 所示。

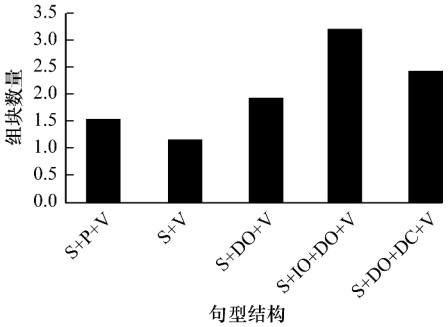

图 3 组块数量平均分布情况

3.2.3 组块生成模板对复述结果的影响

通过对错误藏文复述的结果分析中发现，导致错误的主要原因有 2 个：1）藏文组块生成歧义问题；2）存在特殊分支句型结构。

在组块生成歧义方面，例如“ཤ་ཚ་བའི་ཚོག་ལ་སྦྱན་མོ་མེད།”

“སྐུ་འབུམ་ཐུམས་པ་སྒྲིང་གི་མེ་ཏོག་མཆོད་པ།”,原句分词后组块组合时发生了组合歧义,正确的组合方式为“[ཤ་ཚ་བའི་ཚོག་པ་][སྐུ་རྟེན་མོ་མེད།][སྐུ་འབུམ་ཐུམས་པ་སྒྲིང་གི་མེ་ཏོག་][མཆོད་པ།]”,该句类不能生成复述句,因为谓词部分外只有一个组块元素。例如:“ཀེ་ནི་ཚ་མའི་དབེ་ཚ་རེད།”,该句类不能直接套用藏文组块生成模板,需要特殊处理,例如‘[ཁེ་][ནི་][ཚ་མའི་དབེ་ཚ་][རེད།]”。

以上 2 个错误都归属于 S + P + V 句型结构,该句型结构中存在一些特殊句子,这些分支句型结构对复述句生成实验结果有直接影响。

4　结束语

本文通过语序变换分析藏文简单句型中复述句的生成方法、数量和句型结构,同时对 5 种基本句型结构中每个结构对复述句生成的影响进行实验。实验结果显示,S + P + V、S + V 结构的复述生成数量较少,S + DO + V、S + IO + DO + V 句型结构的复述生成数量较多。此外,当原句中句子成分组块数量较多时,该句的复述生成数量超过几百或几千句,与原句的语义一致。在复述句自动生成过程中,原句的组块分析和生成直接影响复述句语义的准确性。本文方法可以扩展到机器翻译的双语句子对齐、问答系统的答案抽取等应用领域,同时也能解决数据稀疏问题,提高机器翻译评测性能。

复述生成中藏文组块的组合对复述生成结果的影响很大,若组合不合理,复述生成结果与原句子之间的语义就不相等。因此后续将研究句型结构和组块生成方法,进一步提高复述句生成的准确率。

参考文献

[1] BARZILAY R, MCKEOWN K. Extracting paraphrases from a parallel corpus [C] //Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. New York, USA: ACM Press, 2001: 50-57.

[2] GLICKMAN O, DAGAN I. Identifying lexical paraphrases from a single corpus: a case study for verbs [C] //Proceedings of Recent Advantages in Natural Language Processing. Borovets, Bulgaria: [s. n.], 2003: 1-8.

[3] de BEAUGRANDE R, DRESSLER W. Introduction to text linguistics [M]. New York, USA: Longman, 1981.

[4] HALLIDAY M A K, MATTHIESSEN C M I M. An introduction to functional grammar [M]. London, UK: Hodder Education Publishers, 1985.

[5] BARZILAY R, MCKEOWN K R. Extracting paraphrases from a parallel corpus [C] //Proceedings of ACL/EACL 2002. Morristown, USA: Association for Computational Linguistics, 2002: 50-57.

[6] 赵世奇,刘　挺,李　生. 复述技术研究 [J]. 软件学报, 2009, 20(8): 2124-2137.

[7] 王超越. 基于复述技术的汉语情感分析方法的研究 [D]. 哈尔滨: 黑龙江大学, 2014.

[8] ZHAO Shiqi, LIU Ting, LI Sheng. Lexical paraphrasing based on autumatically constructed corpora [J]. Acta Ectrimica Sinica, 2009, 37(5): 975-980.

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

[10] 罗　凌,陈毅东,史晓东,等. 基于复述技术的汉语成语翻译方法研究 [J]. 中文信息学报, 2015, 29(4): 166-174.

[11] IBRAHIM A, KATZ B, LIN J. Extracting structural paraphrases from aligned monolingual corpora [C] //Proceedings of IWP'03. Morristown, USA: Association for Computational Linguistics, 2003: 57-64.

[12] BARZILAY R, LEE L. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment [C] //Proceedings of HLT-NAACL'03. Edmonton, Canada: [s. n.], 2003: 16-23.

[13] DOLAN B, QUIRK C. Brockett C. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources [C] //Proceedings of International Conference on Computational Linguistics. Geneva, Switzerland: [s. n.], 2004: 350-356.

[14] SHINYAMA Y, SEKINE S, SUDO K. Automatic paraphrase acquisition from news articles [C] //Proceedings of the 2nd International Conference on Human Language Technology Research. [S. l.]: Morgan Kaufmann Publishers, 2002: 40-46.

[15] PANG B, KNIGHT K, MARCU D. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences [C] //Proceedings of HLT-NAACL'03. Edmonton, Canada: [s. n.], 2003: 102-109.

[16] ZONG C, ZHANG Y, YAMAMOTO K, et al. Approach to spoken Chinese paraphrasing based on feature extraction [C] //Proceedings of NLPRS'01. Hitotsubashi, Japan: [s. n.], 2001: 551-556.

[17] McKEOWN K R. Paraphrasing using given and new information in a question-answer system [C] //Proceedings of ACL'79. [S. l.]: Association for Computational Linguistics, 1979: 67-72.

[18] TAKAHASHI T, IWAKURA T, IIDA R, et al. KURA: a transfer-based lexico-structural paraphrasing engine [C] //Proceedings of NLPRS'01. Hitotsubashi, Japan: [s. n.], 2001: 37-46.

[19] QUIRK C, BROCKETT C, DOLAN W. Monolingual machine translation for paraphrase generation [C] //Proceedings of EMNLP'04. Barcelona, Spain: [s. n.], 2004: 142-149.

[20] FINCH A, WATANABE T, AKIBA Y, et al. Paraphrasing as machine translation [J]. Journal of Natural Language Processing, 2004, 11(5): 87-111.

[21] 格桑居冕. 论书面藏语单句的两大部类 [J]. 中国藏学, 1994(1): 133-140.

[22] 吉太加. 现代藏文语法通论 [M]. 兰州: 甘肃民族出版社, 2000.

[23] 高定国,扎西加. 藏语单句的基本句型研究 [J]. 中国藏学, 2014(4): 127-131.

[24] 吉太加. 藏文句法研究 [M]. 北京: 中国藏学出版社, 2013.

[25] ABNEY S. Part of speech tagging and partial parsing [M] //CHURCH K, YOUNG S, BLOOTHOOFT G. Corpus-based Methods in Language and Speech Processing. Dordrecht, the Netherlands: Kluwer Academic Publishers, 1996: 119-136.

编辑　金胡考