

英汉机译错误分类及数据统计分析

李梅, 朱锡明

(同济大学外国语学院, 上海 200092)

摘要: 作为一项历时5年探索机器翻译译后编辑自动化研究的重大阶段性成果, 该项目采用汽车技术文献作为研究对象, 通过对10万个英汉翻译句对做机器译文和人工标准译文的对比分析, 发现机译典型错误并进行数据统计分析。介绍该研究所依据的机译错误分类标准, 并以实例加以说明, 得出的机译错误分析数据统计显示, 机译错误率超过八成, 其中词汇类错误占七成。本研究探索的机器翻译译后编辑新思路, 对于降低机器翻译人工成本、提高机译效率和质量具有较高的实际价值和理论意义。

关键词: 英汉机译译后编辑; 错误分类; 数据统计; 汽车技术; 语料库

中图分类号: H 085

文献标志码: A

文章编号: 1009-895X(2013)03-0201-07

Error Patterns and Statistical Analyses of an English-Chinese Machine Translation Corpus

Li Mei, Zhu Ximing

(School of Foreign Languages, Tongji University, Shanghai 200092, China)

Abstract: This paper is based upon a five-year research project on a new attempt at post-editing of machine translation (MT). The research identified the MT error patterns of a text corpus of 100 000 sentence pairs of a car maintenance manual and did statistical analyses of the MT errors. Examples are used in this paper for illustrations of the error patterns. Our statistical analyses reveal that the MT errors reach as high as over 80%, and among the MT errors lexical errors stand at 70%. This research explored a cutting-edge attempt at MT post-editing and was of great significance in improving the MT quality and efficiency as well as lowering the MT cost.

Key words: *post-editing of English-Chinese machine translation; error patterns; statistical analyses; a car maintenance manual; corpus*

机器翻译研究迄今已有50多年历史^[1]。虽然机器翻译已从实验室进入市场, 在人们的工作、学习和生活中发挥越来越重要的作用, 但是机器翻译的质量仍饱受诟病, 甚至出现了所谓的“雪线”现象^[2]。为提高机器译文质量, 多年来人们致力于对机器翻译系统本身的改进^[3-4]。但机器

翻译系统规则结构复杂, 往往牵一发而动全身。修改一项规则虽能解决现存问题, 却又带来新问题。因此, 近年来国内外学者另辟蹊径, 开始研究译后编辑, 即对机器翻译系统从源语到目的语处理后生成的译文进行编辑。译后编辑既可以由人工完成, 一般称为 reviser, 也可以利用软件来实现,

收稿日期: 2013-06-24

基金项目: 教育部人文社会科学研究基金资助项目(07JA740010)

作者简介: 李梅(1963-), 女, 教授。研究方向: 句法学、翻译理论与实践。E-mail: may@tongji.edu.cn

一般称为 post-editor。国外学者对译后编辑的研究有近 10 年历史,主要针对英德、英法互译等^[5]。近年来,国内也开始专注于对译后编辑的研究,而且在译后编辑器软件的设计与开发方面已取得一定进展^[6]。但是,虽有译后编辑器软件的帮助,译后编辑工作的实施仍大都借助系统内置的编辑器^[7-8],由译者在编辑器的提示下实现修改操作。这种方法在一定程度上减轻了译者的负担,但是最大的问题是编辑器不能执行自动修改。由于机器翻译规则的统一性导致同样的错误反复出现,译者需不断修正同样或相似的错误,这不仅使译者不堪重负,而且也大大地降低了速度。

鉴于此,陈欣蓉建议系统化地研究机器译文的错误,以提升机器译文修改研究之效能^[5]。本研究在此基础上提出了译后编辑自动化的研究设想^[9]:在完成语料收集加工后,从分析英汉机器翻译译文的错误入手,通过在词法和句法层面进行语料对比分析,确定具有较高频率和相似度的机译错误典型类型。然后,对其中的典型句法错误进行深度句法分析并做形式化描述。我们设想在不久的将来,将该研究结果用于研制译后编辑自动化软件模块,把该模块加载于相应机器翻译系统的下端,对英汉机译译文进行二次加工,让计算机取代人工来进行译后编辑。这样的译后编辑方式会进一步提高机器翻译的质量,大大加快机器翻译译文人工修改环节速度,提高工作效率。需要指出的是,译后编辑的自动化不能处理所有的机器译文错误,但可以解决那些重复率高、相似度强的错误。据此设想,笔者与同济大学汽车学院合作,将研究对象限定于汽车技术领域翻译^[10],从已建立的约 500 万句的汽车技术翻译文献语料中选取了某品牌汽车维修手册中的近 10 万个英汉翻译句对,采用华建英汉翻译系统进行所选资料的机器翻译,展开了为期 5 年的机器翻译后译文编辑模式研究^[11]。

本文将介绍该项目有关数据处理的过程以及数据统计结果。首先,将探讨该研究所依据的机译错误进行分类,并以实例加以说明。根据此标准对人工标准译文和机器译文进行对比分析后,将机译错误分为词汇、句法和其它三大类,重点介绍机译错误分析结果的数据统计情况。本研究大胆探索机器翻译后编辑新思路,对于降低机器翻译人工成本、提高机译效率和质量具有较高的实际价值和理论意义。

一、机器译文错误分类方式

首先对 10 万个英汉翻译句对(含短语)进行加工分类,在 Excel 文档上将他们分列为英文原文、标准译文。该研究的汉语译文由长期从事汽车专业翻译的专职翻译人员提供,故被视作标准译文,再将英文原文输入机器翻译软件处理形成机器译文,于是便有了英语原文、机器译文、标准译文的平行对比语料库。在展开大规模的机译错误语料分析之前,从 10 万个句对的语料库中抽取了 100 个句对进行机译错误样本分析,同时,参考“中国高校外语专业多语种语料库建设与研究英语语料库”中的错误分类标准^[12],确定了如表 1 所示的机器译文错误分类。

表 1 机器译文错误分类
Tab.1 Categories of MT Errors

词汇类	句法类	其他
词汇术语	句法词序	符号
词汇连词	句法名词短语	标点
词汇词性	句法动词短语	括号
词汇缩写	句法介词短语	物理学单位
词汇漏译	句法被动态	数字
词汇替代	句法不定式	
词汇不译	句法分词	

如表 1 所示,将机器译文语料中的一级错误分成词汇、句法和其他三类。其中在词汇和句法类下分别设 7 个子类的二级错误,在其他类下设 5 个子类的二级错误。现举例说明这三类错误。

1. 词汇类

词汇类的错误共分为 7 个子类,含术语、连词、词性、缩写、漏译、替代和不译等。以下分别举例时为方便读者,凡出错的英语原文以斜体标出。

(1) 词汇术语:指将专门术语误译为一般意义的词汇(不限于名词)。例如:Heater ground 被误译成了“加热器地”,正确的译文应为“加热器搭铁”。

(2) 词汇连词:指将连词如 and 等误译,见例 1。

例 1 英语原文:Check tire type, pressure, and the road surface before making your diagnosis.

机器译文:在诊断之前,检查轮胎类型,压力并且道路表面。

标准译文:在做出诊断之前,检查轮胎类型、压力和路面。

错误类型:词汇连词

上例中,机器译文将连词“and”误译为“并且”,这类错误归为词汇连词错误。

(3)词汇词性:指词性误译,如将名词译成动词。见例2。

例2 英语原文: REAR SEATBACK ADJUSTER ASSEMBLY LH^①

机器译文: 养 育 SEATBACK ADJUSTER 会议 LH

标准译文: 后 排 左 侧 座 椅 靠 背 调 节 器 总 成

错误类型:词汇词性

上例中,REAR 应为名词,意为“后面、后排”,机器译文将其视为动词,故误译为“养育”。请注意:该句机器译文出现多处问题,在此只重点指出相关的词性误译,其它错误暂不讨论。在实际的错误分析中,如例2中的多类错误均一一列出。这也说明了尽管分析的对象语料为10万句对,但是实际错误总数为14万多(详见后面数据分析部分)。

(4)词汇缩写:指原文中的英文缩写词在汽车维修领域有特定含义,而机器将其误译成了其他通用缩写词,见例3。

例3 英语原文: -RR-IG1-1 FUSE

机器译文: - 雷 明 顿 · 兰 德 公 司 - IG1 - 1 保 险 丝

标准译文: -RR-IG1-1 保险丝

错误类型:词汇缩写

该句中RR是汽车维修手册语域中通用的词汇缩写,具有特定的含义,通常不予翻译。

(5)词汇漏译:指原文中的词语在机器翻译中没有译出,见例4。

例4 英语原文: This is the display signal circuit from the multi-display to the television display.

机器译文: 这显示信号巡回从多显示器到电视展示。

标准译文: 这是自多功能显示屏至电视显示屏的显示屏信号电路。

错误类型:词汇漏译

上例中,英语原文中的动词“is”被漏译了,应补充“是”。

(6)词汇替代:指可以翻译成汉语但是在汽车维修手册这一语域中习惯不翻译的专有名词,见例5。

例5 英语原文: Thoroughly mix the repair agent (DuPont paste No.4817).

机器译文: 修理代理人(杜邦粘贴4817号)完全混合。

标准译文: 充分混合维修剂(产品号为4817的DuPont粘接剂)。

错误类型:词汇替代

上例中的“DuPont”虽然翻译成“杜邦”是正确的,但是在汽车维修手册这一语域中通常直接使用英语原文。因此为了尊重行业习惯,保留原文不译。需要指出的是,在做分析时,机器译文对错的取舍不是根据分析员的常识标准,而是严格按照行业标准译文。

(7)词汇不译:指原文中有特定含义的英语,如缩略语不需要译成汉语,而在机器翻译中却被错误地翻译出来,见例6。

例6 英语原文: CAN Communication Line (MS-bus)

机器译文: 罐 装 通 信 线 路 (MS 公共汽车)

标准译文: CAN 通信线路(MS总线)

错误类型:词汇不译

上例中,按汽车维修手册的惯例,“CAN”本不需要译出,而在机器译文中被误译为普通词语“罐装”。

2. 句法类

句法类的错误共分为7个子类,即词序、名词短语、动词短语、介词短语、被动态、不定式以及分词。以下分别举例简要说明。

(1)句法词序:指机器译文中出现的词序错误,如后置、前置等,见例7。

例7 英语原文: The crank position sensor outputs 34 rotation signals per engine revolution.

机器译文: 曲柄位置传感器输出每发动机革命34旋转信号。

标准译文: 发动机每转动一圈,曲轴位置传感器输出34个转角信号。

错误类型:句法词序

上例中机器译文对“每发动机革命34旋转信号”这个时间状语和其所修饰的动词短语“曲轴位置传感器输出34个转角信号”的词序处理有误。

^① 在原始语料中有些英文原文显示为全部字母大写,为保持原语料风格,这里没有做出更改。

(2) 句法名词短语: 指名词短语的误译, 不包括词汇术语中由一个孤立的名词或动词造成的词汇术语误译。这里所牵涉到的是多于1个名词的名词组合, 这样的名词组合所产生的误译是由于句法规则缺失造成。它涉及到的不是单纯的术语问题, 更是对词组组合的方式的错误判断, 故将此类机译错误归为句法错误, 见例8。

例8 英语原文: CHECK BRAKE FLUID LEVEL

机器译文: 检查刹车易流动的水平

标准译文: 检查制动液液位

错误类型: 句法名词短语

上例中, 机器对 BRAKE FLUID LEVEL 这一名词短语的分析有误: 机器将“FLUID”的词性误判为形容词, 从而导致整个名词短语的译文出错。

(3) 句法动词短语: 指动词短语的误译, 不包括词汇术语中的动词性术语误译。与词汇术语不同的是, 这里的误译是由于句法层面的动词中心语理解有误造成, 见例9。

例9 英语原文: Idle speed continues to vary greatly from target speed

机器译文: 无所事事的速度继续从目标速度极大地变化

标准译文: 怠速转速与目标转速持续存在很大偏差

错误类型: 句法动词短语

上例中, 动词短语“continue to vary greatly from target speed”中的核心成分“continue to vary”, 在机译中被介词短语“从目标速度”割裂开来, 造成相关部分的机器译文不知所云。

试比较例10, 动词“放回”有误, 应为“更换”, 这一错误我们归为词汇术语而非动词短语, 因为动词 replace 一般的意义为“放回”, 在汽车维修领域的专业用语则意为“替换”。但这里的误译不涉及词组组合方式问题。

例10 英语原文: REPLACE OUTER REAR VIEW MIRROR ASSEMBLY LH

机器译文: 放回外部后视镜会议 LH

标准译文: 更换左侧车外后视镜总成

错误类型: 词汇术语

(4) 句法介词短语: 指介词短语的误译, 主要指介词单独使用时的误译, 如“I did it for you.”中的 for 归为介词短语。但出现介词与动词搭配的短语误译, 则将其统一归为“句法动词短语”类错误,

如在“I looked after him”中的 after 则应分析为动词短语, 见例11。

例11 英语原文: Front view of wire harness connector: (to Multi-display)

机器译文: 电线马具连接器的正面图: (对多展示来说)

标准译文: 线束连接器前视图: (至多功能显示屏)

错误类型: 句法介词短语

上例中, 英语原文中的介词短语“to…”被机器误译为“对……来说”。

(5) 句法被动态: 指机器译文中的主、被动态的错误, 主要是英语原文的被动句在译成汉语时应还原为主动, 而机器并未这么处理, 还是按照被动句处理, 见例12。

例12 英语原文: When Received Tire Pressure Data is Displayed

机器译文: 当得到的轮胎压力数据是被展示的时

标准译文: 显示接收到的轮胎压力数据时

错误类型: 句法被动态

上例中, 英语原文中的“Data is Displayed”这个被动结构被机器直接翻译成了“数据是被展示的”, 可能是因为机器把“displayed”理解成了一个形容词性的过去分词(类似于“She’s very excited”中的 excited)。即使删除了“是……的”, 剩下的“数据被展示”读起来还是很别扭, 远没有标准译文中转译为主动态意思清晰且不失专业特色。此类误译比例较高, 笔者认为只要在规则上加以约束就可以改善译文质量。

(6) 句法不定式: 指将动词不定式误译, 见例13。

例13 英语原文: Engage the claw as shown in the illustration to set the position.

机器译文: 请爪如确定位置的插图所示。

标准译文: 如图所示, 接合卡爪以设定位置。

错误类型: 句法不定式

上例中, 英语原文中的动词不定式“to set the position”表达的是目的之意, 但被机器误处理成了“illustration”的定语, 所以才会有“确定位置的插图”这样的错误译文。

(7) 句法分词: 指英语原文中的现在分词或过去分词被错误翻译, 见例14。

例14 英语原文: Component without harness

connected; (Blower Motor)

机器译文: 没有马具的零部件连结: (吹风机电动机)

标准译文: 未连接线束的零部件: (鼓风机电动机)

错误类型: 句法分词

上例中,英语原文中的过去分词“connected”从句法上看是修饰“harness”,机器将其处理为动词过去时,与“Component without harness”错误地割裂了开来。

3. 其他类

对于词汇和句法以外的错误归为第三类,包括符号、标点、括号、物理学单位、数字等。由于篇幅关系,这里只列举符号类,见例15。

例15 英语原文: 75° to 105°

机器译文: 75 ? 对 105 吗?

标准译文: 75°至 105°

错误类型: 符号

上例中,机器无法正确识别温度符号“°”而出错。

二、机译错误数据分析

按照第一节中详细说明了英汉机译错误分类标准,参照人工标准译文对近10万个(99 737个)英汉翻译句对中的机译错误进行了对比分析(该对比分析详情参见文献[10])。这里主要介绍根据此分析结果所做的机译错误数据统计。

首先,如表2所示,在99 737句对中有12 939个句对没有出现机译错误(由于篇幅关系这里略去表2中所列的三种正确类型的说明,详情请见参考文献[10]),占句对总数的12.96%。

表2 机器译文无错句统计			
Tab.2 Data Analysis of MT with Few Errors			
编号	类型	数量	所占比例(%)
1	完全正确	3 175	3.18
2	基本正确	3 130	3.13
3	零翻译	6 634	6.65
总计		12 939	12.96

表2统计数据说明,机器翻译错误占所分析句对总数的87%。表3展示了不同类型机译错误出

现的数量及其在错误总数中所占比例。由于同一句中可能有多于一种类型的错误,故表3中误译总计为145 231,超过句对总数99 737。

表3 英汉机译错误分析结果数据统计			
Tab.3 Data analysis of MT with Errors			
编号	子类	数量	所占比例(%)
1	词汇术语	67 428	46.43
2	词汇连词	1 138	0.78
3	词汇词性	7 410	5.10
4	词汇缩写	6 892	4.75
5	词汇漏译	2 844	1.96
6	词汇替代	7 600	5.23
7	词汇不译	9 569	6.59
8	句法词序	13 975	9.62
9	句法名词短语	3 618	2.49
10	句法动词短语	10 470	7.21
11	句法介词短语	5 888	4.05
12	句法被动态	3 836	2.64
13	句法不定式	317	0.22
14	句法分词	880	0.61
15	其他类	3 366	2.32
总计		145 231	100

根据以上数据,将词汇、句法和其他这三类一级错误用饼形图来表示,可以清楚地看到各类错误所占的比重,见图1。

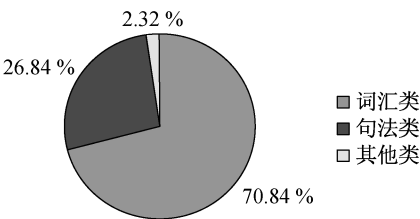


图1 一级错误图示

Fig.1 MT errors at first level

如图1所示,三大类一级错误所占的比例十分悬殊:词汇类错误所占比例高达70.84%,句法类错误占26.84%,其他类错误占2.32%。这些数字告诉我们在该研究所分析的语料中机器翻译的首要问题仍是词汇。如果能在词汇上加以改进,则可以大大提高机译质量。下面来看表2中1~15子类的错误在总数中的占比,见下页图2。

如图2所示,各种错误类型在机器译文中的分布是极不均匀的:从占错误总数的比例来看,比例最高的“词汇术语”类错误,占42%。造成这一现

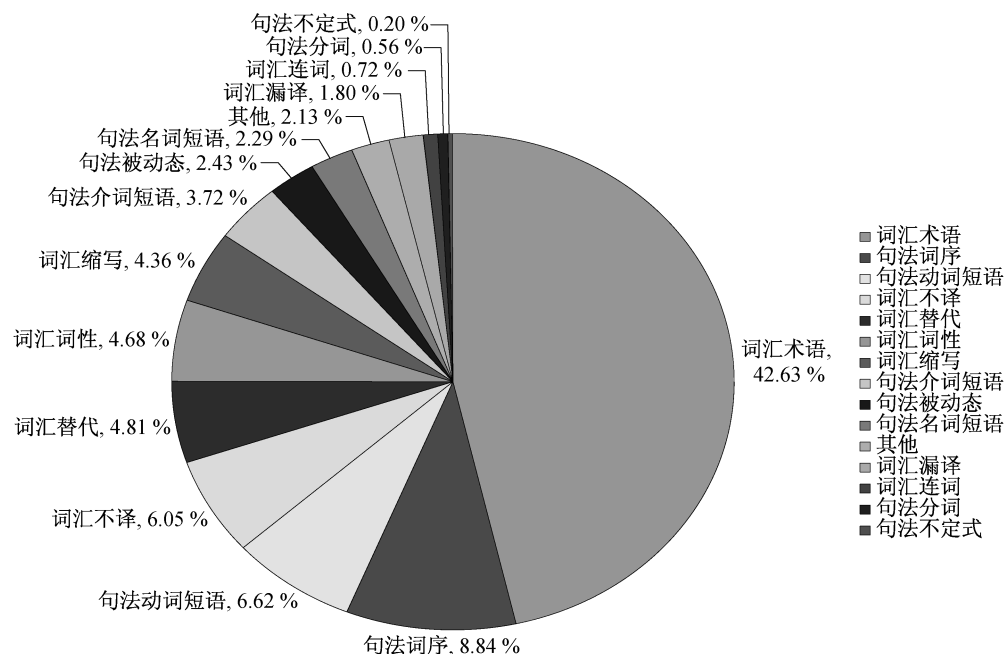


图2 二级错误图示

Fig. 2 MT errors at second level

象的原因是,很多普通的英语单词在特定的技术领域中具有特殊的技术含义,例如: relay block(继电器盒)这个汽车维修领域的术语被翻译成了通用英语意义上的“接替块”。这一事实也从侧面反映出有机翻译的一个突出问题,即科技术语词库的专门化程度不高。上述统计表明,针对某一特定技术领域的术语对译是影响机器译文质量的最大因素。换个角度来看,只要在修改规则中对汽车维修领域中的术语对译进行必要的处理,就可以快速消除机器译文中42%以上的翻译错误,译文的质量无疑将大大提高。

现在来进一步分析词汇类及句法类机译错误。就词汇类错误而言,除比例最高的词汇术语之外,其余6种词汇类错误,即词汇连词、词汇词性、词汇缩写、词汇漏译、词汇替代和词汇不译占错误总数的24.4%,比例也相当高。将词汇层面的各类错误用饼形图表示,可直观地看出各类词汇错误的比重,见图3。

图3中的数据显示,当前的机器翻译仍然未能有效地处理好基本的词汇层面的问题,以至于术语错误的数量竟占了全部词汇错误数量的65%以上。换个角度来看,这一数据也提示研究者,即便在较为简单的词汇层面,只要能投入更多精力进行改进,其实是可以大幅度提高机器翻译的质量的。既然以往的研究表明,越是在受控的领域内,机器翻译越是可以达到较高的翻译质量,典型的如加拿大的TAUM

-METEO 天气预报翻译系统^[13],那么对词汇层面的控制最简单也是最高效的做法就是建立精确的双语对译术语库并加载于原有的机器翻译软件上,则输出的译文质量必会有明显提升。

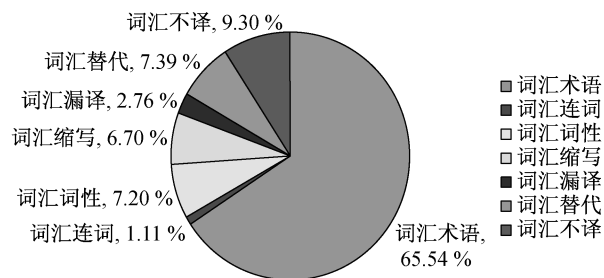


图3 词汇类错误的比例

Fig. 3 Percentage of lexical errors

再来分析句法层面的不同错误,见图4。

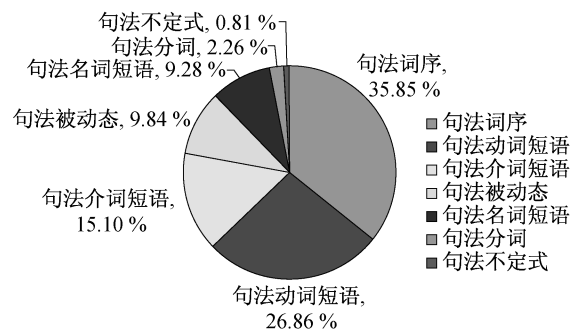


图4 句法类错误的比例

Fig. 4 Percentage of syntactic errors

图4中的数据显示,在句法层面上词序占句法类错误的35.85%、动词短语占26.86%、介词短语占15.10%。据统计提示,如果要在句法层面上尝试通过修改规则来改进机器翻译的质量,重点应当放在“词序”、“动词短语”和“介词短语”这三类上,一旦有所突破,即可消除77%以上的句法类错误。

三、结束语

本研究表明现有机翻译的正确率的确还很低。数据显示,对英语汽车维修手册这一限定语言而言,目前的机器翻译正确率只有12.96%。通过大量的数据分析发现,在机器翻译出现的各类错误中,词汇类错误占比高达70%,其中术语的误译占46%。尽管词汇层面的错误较之句法层面的错误更加容易改正,但却不容忽视。眼下当务之急是对词汇类的各种错误进行进一步分析并使用相应规则来加以纠正。如果能尽快改进词汇层面的机译,正确率会大有改观。此外,研究也揭示,在机器翻译出现的各类错误中,句法类错误的比例占26%以上,所以句法层面的改进虽然困难很大,但仍应是今后机器翻译研究的一个重要内容。本研究的一个重要环节便是进行句法深度分析和形式化描述,也将在生成语言学框架下对句法错误进行深度分析。现有的基于规则的机器翻译系统还有很大的改进余地,而笔者提出译后编辑自动化设想,无需特别复杂的规则和计算就可以显著提升翻译质量,是一个非常值得尝试的。本研究正是沿上述方向迈出的第一步。此后的研究将探讨如何根据本文的数据分析修改规

则,并进行计算机翻译程序转写。希望能与感兴趣的同行们一起努力,为提高英汉机器翻译的质量贡献我们的绵薄之力。

参考文献:

- [1] 冯志伟. 机器翻译研究[M]. 北京: 中国对外翻译出版公司, 2005: 35-64.
- [2] 张克亮. 面向机器翻译的汉英句类及句式转换[M]. 郑州: 河南大学出版社, 2007: vi.
- [3] 计丽丽. 机器翻译系统发展与研究[J]. 中国新技术新产品, 2011(13): 20-21.
- [4] 刘群. 机器翻译技术的发展及其应用[J]. 术语标准化与信息技术, 2002(1): 27-30.
- [5] 陈欣蓉. 评论现行机器翻译译文修饰研究之缺失并且研拟译文修饰研究之具体研究步骤[J]. 国立编译馆馆刊, 2007(3): 97-110.
- [6] 黄河燕, 陈肇雄. 一种智能译后编辑器的设计及其实现算法[J]. 软件学报, 1995, 6(3): 129-135.
- [7] 韩培新. 智能译后编辑器 IPE[D]. 北京: 中国科学院计算技术研究所, 1996.
- [8] 陈震明. 智能机器翻译前后端处理系统的设计和实现[D]. 北京: 中国科学院计算技术研究所, 1998.
- [9] 李梅. Identifying error patterns in MT post-editing[C]. 上海: 第十八届世界翻译大会, 2008.
- [10] 罗季美, 李梅. 机器翻译译文错误分析[J]. 中国翻译, 2012(5): 84-89.
- [11] 李梅, 朱锡明. 译后编辑自动化的英汉机器翻译新探[J]. 中国翻译, 2013(4): 83-87.
- [12] 邹申. 英语专业写作教学语料库建设与研究[M]. 上海: 复旦大学出版社, 2011.
- [13] 袁亦宁. 受控语言: 机器翻译的新途径[J]. 上海科技翻译, 2003(3): 77-80.