中日两国机器翻译研究进展及比较*

□ 张均胜 何彦青 李颖 王惠临 / 中国科学技术信息研究所 北京 100038

摘要：机器翻译研究用计算机实现不同自然语言之间的翻译。自第一台计算机诞生开始，人们一直在研究和探索高质量高效率的机器翻译技术。近年来，基于规则的机器翻译、基于实例的机器翻译和基于统计的机器翻译这几种主要的翻译模式共同存在且相互补充，并有不断融合之势。随着中国和日本在科技、经济和文化交流的不断深入，机器翻译研究对于打破汉语和日语之间的语言壁垒进而推进中日两国各个领域的交流与合作具有重要的应用价值。中国和日本两国机器翻译研究人员已经开展了大量的汉日/日汉机器翻译的理论研究与系统研制，已取得了有效的成果，但距离大规模实际应用和高标准的翻译质量的要求尚有差距。为此，中日两国机器翻译人员有必要在汉日/日汉机器翻译技术与系统研制方面展开合作，特别是在对齐平行文本、实例词典、专业术语词典以及句法分析等共同课题方面展开合作。文章介绍了中日两国机器翻译研究的进展并加以比较，对中日两国在机器翻译领域的合作做了分析与展望。

关键词：机器翻译，基于规则的机器翻译，基于实例的机器翻译，基于统计的机器翻译，系统融合
DOI：10.3772/j.issn.1673—2286.2011.12.004

1 机器翻译简介

机器翻译（Machine Translation），又称为自动翻译，是利用计算机把一种自然语言转变为另一种自然语言的过程。它是自然语言处理（Natural Language Processing）的一个分支，与计算语言学（Computational Linguistics）、自然语言理解（Natural Language Understanding）之间存在着密不可分的关系。被翻译的语言称为源语言，翻译成的结果对应的语言称为目标语言。

多年来，机器翻译一直被认为是信息社会对计算机技术最具挑战性的研究课题之一。机器翻译研究涉及人工智能、数学、语言学、计算语言学和语音技术等多种学科与技术，是一项综合性研究课题。其中，语言学家提供适合于计算机进行加工的词典和语法规则；数学家把语言学家提供的材料形式化和代码化；计算机科学家给机器翻译提供软件手段和硬件设备，并进行程序设计。机器翻译效果的好坏取决于上述几个方面的共同努力。

很多人对机器翻译有些误解，认为机器翻译结果与人工翻译偏差大，不能帮人们解决任何问题。由于机器翻译运用语言学原理，机器自动识别语法，调用存储的词库，自动进行对应翻译，但是因语法、词法、句法发生变化或者不规则，造成机器翻译结果出现错误是在所难免的。

机器翻译的目的是辅助人工翻译，为人工翻译减轻负担和提高效率，在部分场景和任务下替代人工，而决非要彻底取代人工翻译。想要计算机完全替代人工完成一切翻译任务的企图，恐怕是一个永远都不可能实现的梦想。在上个世纪人们早就已经认识到，在机器翻译研究中实现人机共生、人机互助比

* 本文受中国科学技术信息研究所学科建设“自然语言处理”课题（XK2011—6）、中国科学技术信息研究所重点工作“多语言信息获取关键技术研究与应用示范”课题（ZD2011—3—3）、中国科学技术信息研究所科研项目预研资金（YY—201122）和国家社科基金项目“基于本体的跨语言信息检索理论与实验研究”（06BTQ030）资助。

追求全自动的高质量翻译更现实。针对日益剧增的多语种信息，让计算机粗略地处理一遍，经筛选之后，如果需要再由人工或者采用机助人译的方式完成细加工过程，被认为是一种可行的处理办法。翻译中的“信、达、雅”永远都是人们孜孜以求的目标，但是在人类对于自身大脑翻译的思维过程都还没有弄清楚以前，要求计算机高质量地自动翻译，甚至翻译小说、散文、成语或诗歌等文学作品是不现实的，而且在许多情况下人类自己都做不到。因此，人工翻译与机器翻译系统之间应该是互补、互助的关系，而不是相互竞争[5]。

机器翻译技术应用需求及市场巨大。随着经济全球化时代的到来，如何克服语言障碍已经成为国际社会共同面对的问题。美国Allied Business Intelligence（ABI）公司曾经对世界翻译市场做过调查，结果显示到2005年时，翻译市场规模已达220多亿美元。在欧盟委员会中，各机构每年的翻译费用达10亿多欧元，对于20种官方语言，每种语言每天需要80名口译人员。有关调查表明，目前中国有100多万翻译人员，其中科技翻译人员就达40多万人，国内翻译年产值大约为60亿元人民币。

机器翻译技术对情报部门具有重大意义。日益激增的多语种政治、经济、军事等情报信息已使情报部门不堪重负，单单靠人工翻译和阅读变得非常困难。为此，美国国防预先研究计划局（DARPA）和欧盟第六框架等都已投入巨资开展该项技术研究。由此可见，机器翻译具有极其广阔的应用前景。从某种程度上讲，多年来机器翻译始终是国际学界、商界甚至军界共同角逐的必争之地。

从目前国际研究现状来看，机器翻译的若干理论问题一直没有从根本上得到解决，许多方法和技术有待于进一步研究和探索。机器翻译系统的性能也确实不尽如人意，无论是系统译文的质量，还是系统的自学习能力、知识库维护和更新能力，以及对各种非规范语言现象的处理能力等，都还有待于大幅度提高。中国数学家、语言学家周海中曾在论文《机器翻译五十年》中指出：要提高机译的质量，首先要解决的是语言本身的问题，而不是程序设计问题；单靠若干程序来做机译系统，肯定是无法提高机译质量的[4]。另一方面，机器翻译已经在某些限定领域为人们提供了快捷方便的翻译服务，例如，天气预报翻译、产品说明书翻译等等。即使在无领域限制的网页在线翻译等方面，有些软件也在一定程度上提供了便利，而且计算机辅助翻译和译后编辑（post-editing）功能都为人工翻译提供了一定的帮助。总之，机器翻译既不像有些人批评的那样一无是处，又不像有些人吹捧的那么完美无缺。但不可否认的是，机器翻译作为一个科学问题在被学术界不断深入研究的同时，企业家们已经利用它从市场上获得了丰厚的利润。

日本科学技术振兴机构高级执行理事川上伸昭发言

2 机器翻译研究简史

机器翻译研究始于20世纪50年代。1954年，美国乔治敦大学（Georgetown University）在IBM公司协同下，用IBM-701计算机首次完成了英俄机器翻译试验，向公众和科学界展示了机器翻译的可行性，从而拉开了机器翻译研究的序幕。当时美国致力于开发英俄翻译系统，而日本则重点开发日英/英日翻译系统。中国机器翻译研究开始于1956年，并在1959年成功地进行了中国首次机器翻译试验表演。中国机器翻译研究从一开始就得到了国家的高度重视。早在1956年，机器翻译就被列入了当时的《科学发展纲要》，以后则列为“六五”、“七五”，以及“863”等

重大科研项目。中国的机器翻译研究从一开始就具有多单位、多方面不同知识结构人员的协同攻关的特点。这是这项研究自身的特点所决定的，它需要至少计算机科学、数学、语言学等多方面的知识[3]。

为了对机器翻译的研究进展作出评价，1964年美国科学院成立了语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，简称ALPAC委员会），开始了为期两年的综合调查分析和测试。1966年11月，该委员会公布了一个题为《语言与机器》的报告（简称ALPAC报告），该报告全面否定了机器翻译的可行性，并建议停止对机器翻译项目的资金支持。这一报告使得机器翻译研究陷入了近乎停滞的僵局。无独有偶，在此期间，中国爆发了十年“文革”，基本上中国的机器翻译研究也停滞了。机器翻译研究整体步入萧条期。

70年代中期，中国机器翻译研究从停滞走向了复苏，是协同攻关的特点体现最充分的时期。当时在中国科技情报所（中国科学技术信息研究所前身）的组织下集中了许多部委的研究人员在社科院语言所的专家的具体指导下协同攻关。在这一时期，还开始向国外派出人员学习和引进技术，并与当时已在国际享有盛名的机器翻译研究机构进行了交流。同时，社科院语言所开始培养机器翻译专业研究生。

20世纪80年代，欧洲机器翻译的研究重心转向欧共体多国语言之间的多语言翻译。日本则把研制日英/英日翻译的实用系统作为课题，以JICST（其后成为JST的一部分）和当时任职于京都大学的长尾真教授为中心，并于1986年研制了科学技术论文的日英/英日机械翻译系统，其改良版也在之后得到广泛应用[2]。同时，中国的机器翻译研究产生了两个在中国机器翻译史上具有重要意义的实用化系统——军事科学院研制的“KY-1”英汉机器翻译系统，它获得了国家科技进步二等奖，后来被开发为“译星”，成为中国第一个商品化系统；中科院计算所研制的“863-IMT”英汉机器翻译系统，获得了国家科技进步一等奖，该系统带来了十分可观的效益[3]。另外，由邮电科研院研制的“MT-IR-EC”是一个非常实用的通讯题录系统，使刊物的发行效率得到很大的提高，并因此成为了第一个荣获国家科技进步奖的机器翻译系统。

日本国立国会图书馆馆长长尾真发言

20世纪80年代，中国参加了由日本发起的亚洲五国机器翻译研发的合作项目。国内近10个单位参加了这一长达7年的国际项目。这次的大协作对于培养人才、传播技术、积累资源（如词典等）以及使中国的机器翻译研究走向世界，都有着深远的影响。这个时期又正值国内“七五”计划，给了更多的单位和研究人员参与机器翻译研究的机会。在此期间，清华大学和南京大学研制了实用的日汉机器翻译系统。中国科大在机器翻译通用工具方面进行了富有成果的研究。北京大学研制成功了机器翻译系统自动评估系统，这在国内外尚属首例。

20世纪90年代初期至今，统计机器翻译逐步走上历史舞台，成为机器翻译研究的热点。同时，中国的机器翻译也走入了快速发展的时期，出现了许多商品化系统。近期的机器翻译系统大体上有这样一些特点：多数配有大规模的多种领域的专业词典，多数能在网上运行，有相当不错的方便用户的界面。新的应用领域的机器翻译研究，如对话翻译系统的研发等也已开始。

3 机器翻译方式

机器翻译经过半个多世纪的发展和研究，各种翻译方法不断涌现，其中基于规则的翻译方法和基于语料库的翻译方法是两个主要的

流派。

在20世纪90年代之前，基于规则的方法在机器翻译中占据主导地位。基于规则的方法将翻译过程分三个阶段：对输入文本进行分析，形成源语言抽象的内部表达；将源语言抽象的内部表达转换成目标语言的内部表达；根据目标语言的内部表达生成目标语言文本。该方法的优点在于可以较好地保持原文的结构，对于语言现象已知或者句法结构规范的源语言句子翻译效果尤为理想。然而，它的人工规则编写成本高昂、规则一致性难以保证、相关知识库的建立和维护也很困难，这些都使得该方法无法适应大规模数据发展的需要，很难逃脱被取代的命运。

到20世纪80年代中后期，基于语料库的机器翻译方法获得长足发展，逐渐占据了主导地位。这种方法建立在大规模收集互为译文的双语语料的基础之上，有两个主要分支：基于实例的机器翻译（Example-based MT，EBMT）方法[6]和基于噪音信道模型的统计机器翻译方法[1]。

基于实例的翻译方法主要通过在双语语料库中查找最为相似的翻译实例来获得目标翻译。该方法的优点是译文自语料库中的实例直接变换而来，译文质量较高，速度快，能够有效地解决一些基于规则的机器翻译系统难于处理的问题。其主要缺点在于覆盖率方面，基于实例的翻译往往难以达到很高的覆盖率，对齐正确率直接影响到翻译的正确率。通过实例的组合得到目标译句可以借助多种做法，如基于字符的匹配，基于短语结构的匹配，基于依存关系的匹配等。基于实例的机器翻译引擎通常作为其他翻译引擎的补充，而很少作为独立的机器翻译引擎使用。

统计机器翻译（Statistical MT，SMT）建立在噪音信道模型之上，它先要对翻译过程建立数学模型，利用双语语料库估计模型参数，进而根据模型及估计的参数执行翻译。统计机器翻译是当前机器翻译领域的研究热点，在美国国家标准和技术研究所（National Institute of Standards and Technology，NIST）信息部语音组主导的机器翻译国际评测中，从2002年以来，统计机器翻译系统的成绩一直名列前茅。这说明统计机器翻译系统的性能已超过了基于规则的和其他类型的翻译系统，成为机器翻译的主流方法。统计方法的优势在于可以利用大规模的语料，不需要太多的人工参与，实现相对简单。近几年来，基于统计的翻译方法得到了充分发展，不断地推动着机器翻译水平的提高。

根据翻译模型的不同，统计机器翻译可以分为基于词的翻译模型、基于短语的翻译模型和基于句法的翻译模型。

3.1 基于词的统计翻译模型

在基于词的统计机器翻译模型中，翻译过程被看成是一个信源信道模型[1]，将词对齐作为隐变量引入翻译过程，利用EM算法[27]从句子级对齐的语料中自动训练出单词对齐，利用与词对齐相关的概率，通过动态规划算法搜索出对应源语言句子的最优目标语言句子。该模型对从词到词的自动生成过程进行建模，数学建模描述十分严密。它所使用的统计方法功能强大。但是，基于词的模型翻译单元是单个的词，只能学习到由词到词的翻译知识，对词的上下文未作考虑，不适合翻译习惯表达、成语等结合紧密的源语言串；它的重排序能力较差；模型比较复杂，需要估计的参数太多。

3.2 基于短语的统计翻译模型

针对基于词的统计翻译模型存在的缺陷，人们提出了基于短语的统计翻译模型。基于短语的统计机器翻译方法[28-31,33,53]将“双语短语”定义为源语言或者目标语言任意连续单词组成的互为翻译的串对。它的主要翻译过程包括：根据均匀分布的假设将源语言句子划分成短语，利用预先抽取的短语翻译对表将每一个源语言短语翻译成目标语言短语，然后利用重排序模型对目标语言短语进行重排序，最终得到目标语言句子。

基于短语的统计翻译模型的优势在于它利用短语作为单位，很好地学习到了词的上下文信息和局部重排序信息，长于翻译习惯用语和多语表达。因而基于短语的统计机器翻译改进了基于词的统计翻译方法，成为统计机器翻译的新研究热点。但是由于受到翻译模型本身的限制，基于短语的翻译模型又存在难以克服的缺陷。这主要表现在下面三个方面：

（1）短语的重排序能力差：已有的重排序模型[35-40,42,43]大都难以做到全局的重排序，只停留在语言表面，只进行简单的短语匹配和位置调整，没有涉及更为深层的语言学知识。

（2）短语表的构建鲁棒性差：在现有的基于短语的统计机器翻译

模型中，短语要求完全匹配，只要有一个字或者词不一样，这个短语对就不能使用，甚至可能导致这个源语言短语无法翻译。

（3）短语的泛化能力差：基于短语的翻译系统中的短语都是连续的，在源语言端和目标语言端，都呈现为连续的词串。在翻译中无法使用非连续短语是该模型的短语泛化能力差的主要表现。

以上三个问题正是基于短语的统计机器翻译模型自身不能解决的问题。针对这些缺陷，研究人员试图引入更深层次的语言结构和句法信息来改善统计机器翻译的性能，因而出现了基于句法的统计机器翻译模型。

3.3 基于句法的统计机器翻译模型

基于句法的统计翻译模型的特点在于引入语言的结构信息，利用其层次化重排序能力、泛化能力和处理非连续短语的能力生成目标翻译，较之基于短语的统计翻译模型，因其借鉴了基于规则的翻译方法的经验，又结合了基于语料库的翻译方法的精华，顺应了机器翻译研究发展的必然趋势。

按照句法信息是否借用语言学知识，基于句法的统计翻译模型可以分为两类：基于形式句法的统计翻译模型（formally syntax-based statistical machine translation models）和基于语言学句法的统计翻译模型（linguistically syntax-based statistical machine translation models）。

基于形式句法的统计翻译模型[17-19,51]使用形式化的结构表示了句子的某种层次性划分，它的层次化结构能够实现全局的短语重排序，使用非终结符来标识短语，使得该翻译模型可以使用非连续短语，从而能够改善基于短语的统计机器翻译的缺陷。但是因为各个节点和节点的关系不具有语言学的意义，没有利用深层次的语言学的知识，加上同步文法要求源语言树和目标语言树是同构的，因而形式化结构的表达能力有限。

基于语言学句法的统计翻译模型使用具有语言学意义的层次结构，其节点本身和节点之间都使用了语言学知识，常常利用源语言端或者目标语言端或者两端的句法分析树。现有的句法分析树又包括短语结构树和依存树。前者描述了句子的组成成分以及各成分之间的关系，体现了句子的句法结构；后者描述了词与词之间的关系，反映了更多的语义知识。基于短语结构树的统计机器翻译模型，根据句法分析树所在的语言端，又分为树到串翻译模型（tree-to-string translation model）、串到树翻译模型（string-to-tree translation model）和树到树翻译模型（tree-to-tree translation model）。串到树翻译模型[34,40,41]的基本思想是假设目标语言端的树经过噪音信道后被异化为源语言的串，通过解码将源语言的串还原成目标语言的树。树到串的翻译模型[20,42]利用概率化的规则，转化的是从源语言的树到目标语言的串。树到树的翻译模型[43-47]通过树转换或者同步分析试图实现从源语言树到目标语言树的生成。与树到串翻译模型和串到树翻译模型相比较，树到树的翻译模型显然更复杂，需要解决语言间的结构性差异问题。这类方法主要研究源语言子树到目标语言子树的映射。由于树到树翻译模型面临更多句法分析技术问题和数据稀疏问题的挑战，因而至今没有达到理想的效果。较之短语结构树，依存结构树是词汇化的，它能够体现出词汇间更多的语义关系，减小树结构的差异性[48]。因而，越来越多的研究者采用依存句法树信息来进行机器翻译[49-51]。

综上所述，基于句法的统计机器翻译系统使用层次结构弥补了基于短语的统计机器翻译在短语的连续性和重排序模型上的缺陷，提高了翻译质量。但是截止到目前，基于句法的统计机器翻译仍然没有能够大范围流行起来，一些现存的相关技术问题限制了该类模型的发展。其一，翻译模式如果与大规模数据相结合，解码器就必然面临庞大的计算压力和存储负担。其二，高质量鲁棒性强的句法信息的获取还很难保证，尤其是针对汉语。其三，语言本身的差异性导致双语句法树的异构，这是句法模型本身难以克服的。

针对上述问题，需要寻找比基于句法的统计机器翻译方法更为先进的翻译模式。按照统计机器翻译著名的金字塔图[52]，研究者已经开始了使用语义信息来进行机器翻译研究。这方面的工作目前还集中在使用词义消歧和语义角色标注来改善统计机器翻译[12-15]。

在众多的机器翻译方法中，很难说哪一种翻译模型在翻译效果上具有绝对的优势。因此人们也采取了另外一种途径来改善机器翻译性能，即多机器翻译系统融合。这项技术开始于1994年[16]，随着机器翻译的应用需求不断增加，越来越多的机器翻译系统融合方法不断涌现。

3.4 多机器翻译系统融合方法

常用的多机器翻译系统融合可以从句子、短语和词三个级别上独立进行，同时也可以将三种方式结合起来进行融合。句子级的融合策略[15]实际上是一种句子重排序，使用单机器翻译引擎所使用的特征以外的新特征的信息从合并的N-best列表中选择得分最高的翻译作为融合结果。短语级的多系统融合[15]方法是利用翻译候选和源语言句子的对齐重新抽取短语组成短语表，然后再进行重新解码。解码模型同标准的短语翻译系统一样。目前国内外主要的多系统融合策略大都是词语级系统融合[7-11]，其核心思想是使用最小贝叶斯风险（Minimum Bayes Risk，MBR）解码器从所有系统的N-best结果中选择一个最优结果作为对齐参考，然后将其余翻译结果与该对齐参考进行词对齐构建混淆网络，然后再利用投票策略重组识别结果，产生一个得分最高的路径。该方法利用词作为单位，保证了混淆网络最大的候选空间，使得新生成的翻译假设可能性增大，但是也增加了翻译结果变差的可能性。

通过以上分析可以看出，基于规则的机器翻译、基于实例的机器翻译和基于统计的机器翻译各有其优缺点。笔者认为，今后的机器翻译发展一方面会在理论模型探索上继续向深层语言学发展，另一方面在实际应用中将走向融合基于规则的机器翻译、基于实例的机器翻译以及基于统计的机器翻译三者优点的混合性系统。

4 中日机器翻译研究

中日两国在科技研究领域、工业技术以及经济文化交流方面交流越来越多，越来越深入。近代中国科技水平相对落后，中国科研人员大量地阅读日文科技文献。然而进入21世纪以来，随着中国国力及其在世界舞台的影响日渐增强，汉语热正在世界范围内悄悄兴起。中国的科技发展引人瞩目，日本研究人员也开始大量阅读中国的科技文献。日本制定了第4期科技基本计划，将在今后的5年内投入25万亿日元研究日汉/汉日机器翻译。在科技领域，中日两国已经取得了显著的研究成果，并有望在将来取得更大的研究成果。然而，中日两国的科技研究成果基本上都以本国语言各自发表，难以在两国之间简便地投入应用，该问题正日趋严峻，严重阻碍了中日两国科技信息共享和交流。另外，随着中国科技与工业的迅猛发展，中日贸易规模也在扩大，时至今日中国已经成为日本最大的贸易伙伴国。为此，中日两国很有必要开发日汉/汉日机器翻译系统。如果日汉/汉日机器翻译系统能够走向实际运用，将有利于中日两国在工业技术与科技领域不断加深了解和合作。

一直以来，中日两国机器翻译研究人员在机器翻译技术研究与应用方面不断展开交流与合作。长尾真教授在京都大学长年从事机器翻译的研究，也与中国机器翻译的研究人员进行了亲密的交流。他自2007年担任国立国会图书馆长以来，一直呼吁中国国家图书馆、韩国国立中央图书馆，建议中日韩三国国立图书馆运用机器翻译系统为相互间的资料利用开拓道路。2010年夏天，中日韩国立图书馆正式签订合作协议。日本国会图书馆首先导入了翻译效率较高的日韩/韩日翻译系统，在日韩两国的国立图书馆之间开辟了相互利用资料之路。目前正在导入中日机器翻译系统进行测试，希望能与中国国家图书馆、中国国家科技文献中心等文献服务机构之间实现资料的相互利用，只是机器翻译系统的完善还需要时间，在不远的将来肯定能够开辟资料相互利用之路。中国国家科技文献中心是收藏中国科技相关信息的代表性权威机构，如果能够与日本的国立国会图书馆之间通过机器翻译系统实现资料的相互利用，对于中日两国的科技信息交流和资源共享将起到巨大的促进作用。

在过去的五年中，中国的十一五国家支撑计划项目“多语言信息服务环境关键技术研究与应用”，旨在研究和集成开发英汉计算机辅助翻译系统，包括英汉网络在线翻译系统和面向专业翻译人员的英汉机助翻译平台，研究开发英汉跨语言检索查询接口系统，用于实现文献信息系统的英汉翻译和双语言检索服务，为多语言信息服务技术的全面推广和应用以及形成产业化产品奠定基础。该项目已经顺利结题，目前在改进现有英汉机器翻译引擎质量的同时，着手准备将日汉/汉日机器翻译系统应用到国家科技文献中心网络系统中。

日汉/汉日机器翻译与日英/英日机器翻译相比难度更大。原因在于中文是一种属于孤立语的语言，语法结构不同，没有活用的形与时态，句子不能以单词为单位进行切分，而是一串连续的汉字，这些特点与欧美语言以及日语全然不同。因此，很难以单词为单位对中文语句

进行切分，而且中文句法结构分析也不简单。

为了继续推进日汉和汉日机器翻译的发展，迫切需要整合中日两国的机器翻译研究成果、各种精确的词典、对齐的平行文本数据等，建立实用化的汉日和日汉的机器翻译系统。这样高质量鲁棒性强的机器翻译系统，不但会增进机器翻译研究本身的发展，将其应用于图书馆之间的跨语言信息服务，或者是经济、文化交流中的实际场景，对于中日两国的各方面信息交流和资源共享都会起到巨大的促进作用。

4.1 国内的机器翻译研究现状

受国际环境的影响，我国的机器翻译也经历了初始期、停滞期、复苏期，在1987年迎来了繁荣期。之前，我国的机器翻译研究的每一个时期都比国外机器翻译的同样时期稍微滞后。但是，近年来，尤其是21世纪初的前10年间，我国机器翻译学者在国际学术界的影响力在不断提高，已有多人担任过ACL（Annual Meeting of Association for Computational Linguistics）、COLING（International Conference on Computational Linguistics）、EMNLP（Conference on Empirical Methods in Natural Language Processing）、IJCNLP（International Joint Conference on Natural Language Processing）等机器翻译领域主席，每年都有多篇ACL等顶级会议论文发表，这标志着我国的机器翻译研究已经在国际机器翻译领域崭露头角。

目前，基于语料库的机器翻译仍然是机器翻译的主流方法，这种建立在大规模真实文本处理基础上的翻译方法仍然是当前机器翻译的总体特征。中国的机器翻译研究在理论探索方面，遵循着统计机器翻译方法中从词到短语进而句法的发展脉络，不断地向前拓进着。在基于句法的统计机器翻译模型研究中，中国科学院计算所取得了一些突出的研究成果。论文[27]在括号转录文法（Bracket Transduction Grammar，BTG）[25]的基础上，使用最大熵模型预测任意相邻成分保序或者逆序的概率，该模型为基于形式句法的统计机器翻译中的经典性工作。利用源语言树到目标串的对齐模板作为规则，自底向上地遍历源语言结构树中每个节点，搜索与之匹配的对齐模板，最终生成目标译文[28]。该树到串的翻译模型为基于句法的统计机器翻译中有代表性的树到串翻译模型，并获得了COLING/ACL 2006年的“Meritorious Asian NLP Paper Award”。之后，论文[21]提出了基于句法森林的翻译模型，使用句法森林来代替1-best的句法树，很好地缓解了句法分析错误对翻译规则集合质量以及最终翻译质量的影响。该模型属于基于句法的统计机器翻译的前沿性工作。另外，计算所在联合句法分析和机器翻译方面也作了尝试[22]。

在机器翻译的系统实现方面，我们以全国机器翻译研讨会（China Workshop on Machine Translation，CWMT）所举办的机器翻译评测的情况为例，可以一窥国内机器翻译系统实现的现状①。

CWMT由中文信息学会主办，每年举办一次。该研讨会最具特色的地方就是其中的机器翻译系统评测活动，由中国科学院计算技术研究所组织，所有参评系统在相同的时间段和相同的测试集上运行，并采用相同的评判标准和指标对系统运行结果进行打分和排名。自2007年以来已经成功举办4次的CWMT系统评测是对该领域过去一年内技术进展状况的一次检阅，是公平、公开的技术竞争，具有很强的说服力。表1列出了历年来CWMT组织的机器翻译评测设置的项目。在最初评测只集中在汉英和英汉的新闻领域的项目上，之后将翻译领域拓展到科技领域上，在汉英新闻领域加入了系统融合的评测，这说明国内机器翻译研究主要集中在英汉和汉英的翻译方向上。在2011年最为显著的特点是加入了日汉新闻领域的评测项目和少数民族语言与汉语的翻译项目，包括蒙汉日常用语、藏汉政府文献、维汉新闻领域、哈汉新闻领域、柯汉新闻领域。参与日汉翻译的国内单位只有6家。这一方面说明机器翻译的需求性很大，不只表现在不同国家的语言之间，也实际存在于我国多民族的不同语言之间；另一方面也说明机器翻译的研究者正在将实验室的研究成果初步向实际应用推广。

表2列出了CWMT2011的国内参评单位，此次评测共有19个单位报名参加，其中国内单位15家，国外

① 文中表格参考了历次全国机器翻译研讨会的评测报告。
② http://www.statmt.org/moses/.
③ http://www.nlplab.com/NiuPlan/NiuTrans.html.

表1 参评项目

年份	汉英新闻	英汉新闻	英汉科技	日汉新闻	蒙汉日常用语	汉蒙日常用语	藏汉政府文献	维汉新闻	哈汉新闻	柯汉新闻	汉英新闻系统融合
2007	√	√									
2008	√	√	√								√
2009	√	√	√			√					√
2011	√	√	√	√	√		√	√	√	√	

表2 CWMT2011参评单位及参加项目情况

报名单位	汉英新闻	英汉新闻	英汉科技	日汉新闻	蒙汉日常用语	藏汉政府文献	维汉新闻	哈汉新闻	柯汉新闻	总计
北京航空航天大学			√	√						2
北京交通大学	√	√	√							3
南京大学	√	√	√	√						4
西安理工大学	√		√							2
中国科技信息研究所			√							1
中科院计算所	√	√	√	√	√	√	√	√	√	9
中科院自动化所	√	√	√	√	√	√	√	√	√	9
东北大学	√	√	√	√	√	√	√	√	√	9
中科院软件所	√					√				2
厦门大学	√	√				√				3
哈尔滨工业大学	√	√	√	√		√				5
内蒙古师范大学					√					1
新疆大学							√	√	√	3
中科院合肥物质科学研究院智能机械研究所							√			1
中科院新疆理化技术研究所							√			1

表3 CWMT2011每个项目上自动评测第一名的系统情况

系统	汉英新闻	英汉新闻 Progress	英汉新闻 Current	英汉科技	日汉新闻	蒙汉日常用语	藏汉政府文献	维汉新闻	哈汉新闻	柯汉新闻
单系统	√	√	√				√			√
融合系统				√	√	√		√	√	

单位4家。从各个单位的分布范围，可以看出国内机器翻译研究主要集中在中国科学院的研究所、知名的大学以及公司。当然，中国还有一些从事机器翻译研究的单位没有参加此次评测。从每个单位参与项目的

个数上可以一窥其实力，其中中科院自动化所、计算所和东北大学的实力比较突出。

表3列出了CWMT2011每个项目上自动评测第一名的系统情况，可以看出机器翻译单一系统和多机器翻译系统融合平分秋色。所有的参评系统都使用了基于统计的翻译方法，中国科学技术信息研究所将规则方法和统计方法进行了系统融合[26]。评测中翻译表现比较突出的是中科院自动化所的基于功能的统计机器翻译系统RoleTrans[23]，这个翻译系统集短语翻译、MEBTG和非连续短语为一体。语言模型使用了前向语言模型和后向语言模型。很多参评单位仍然使用了Moses②系统，这是一个完整的基于短语的统计翻译框架，是机器翻译发展的一个里程碑，为后续研究者研究新的翻译模型和比较翻译效果奠定了坚实的基础，大大推动了机器翻译技术的发展。值得一提的是东北大学的NieTrans③系统，该系统已经开放源代码，希望该系统能够成为继Moses之后又一个推动机器翻译发展的工具。在日汉机器翻译方面，南京大学采用了基于规则和日语格语法语义分析的转换翻译系统[24]。唯一的一家使用基于实例的机器翻译方法参评的是北京航空航天大学[25]。

虽然规则系统在CWMT的评测中并不多见，但是在各个公司却使用很普遍，国内的CCID、华建和中软都开发了规则系统。富士通公司在进行专利翻译时也使用了规则系统，其将统计方法用来进行预处理再用于规则系统的方法，以及使用规则方法将复杂句转变为简单句再使用统计系统的方法都值得业界借鉴。

4.2 日本的机器翻译研究现状

在日本，人们希望机器翻译这项战略性技术能够发挥重要的作用，帮助日本不断提升其在世界经济中的地位。

在美国乔治敦大学进行了世界上第一次机器翻译试验之后，日本于1958年开展了机器翻译的试验。在日本的一些主要的大学和政府或者工业联盟都有机器翻译和自然语言处理研究组。日本机器翻译专家长尾真最早于1984年提出基于实例的机器翻译，他在《采用类比原则进行日-英机器翻译的一个框架》一文中，探讨了由实例引导推理的机器翻译方法，至今仍是机器翻译的经典性方法之一。基于该思想，长尾真和佐藤（S. Sato）还开发了基于实例的机器翻译系统MBT1和MBT系统。20世纪90年代初期，日本学者北野（Kitano）在京都大学期间，使用大规模并行计算，采用基于实例的方法进行语音翻译试验，证明了毫秒级的实施口语语音翻译是可实现的，他也因此得到了1993年度IJCAI的计算机与思维奖。目前，基于实例的机器翻译方法在日本得到了进一步的发展，研究者使用词典和句法信息来改进此类翻译引擎。

几个日本最大的工业公司正在开发机器翻译系统，很多机器翻译系统已经商业化。日本的很多实用化的机器翻译系统为规则系统，主要集中在日语和英语两个语言之间，比如日本富士通公司开发的ATLAS系统、日立公司开发的HICATS机器翻译系统、日本电器公司开发的PIVOT机器翻译系统、三菱电机公司开发的MELTRAN机器翻译系统、冲电气公司开发的PENSEE机器翻译系统、理光公司开发的RMT机器翻译系统、三洋电器公司开发的SWP-7800机器翻译系统、东芝公司开发的TAURAS机器翻译系统、日本布拉维斯公司研制的BRAVICE PAK 11/73机器翻译系统、日本夏普公司开发的Power机器翻译系统等。

从1997年，日本科学促进协会（Japan Society for Promotion of Science，JSPS）和国家信息机构（National Institute of Informatics，NII）等几家科研机构合作组织了NTCIR（NII Test Collection for IR Systems）计划，这是一系列的评测活动，用于促进信息获取技术的研究，包括信息检索、问答、文本摘要、文本抽取等，其中包括专利信息的机器翻译评测。在2011年NTCIR' 9的评测中，专利翻译的任务包括了汉英、日英和英日三个翻译项目，共有来自不同国家的21个研究组参与了评测，18个参与汉英，12个参与日英，9个参与英日。由于此次评测的评测结果还没有公布，无法得知参评单位有哪些，但是从NTCIR专门设置专利机器翻译项目的评测，可以看出日本对于机器翻译研究的战略性地位的重视程度。

4.3 中日两国机器翻译研究的不同

对比中国和日本两国的机器翻译研究现状，主要的不同表现在以下几点：

（1）语言对不同：中国对汉英和英汉的研究比较多，日本着重对英日和日英的机器翻译研究。这一方面原因在于英语是当今世界上主

要的国际通用语言之一，也是世界上最广泛使用的语言，其使用范围之广，涉及国际政治、军事、经济、科技、文化、贸易、交通运输等多领域。另一方面也受国际经济政治局势的影响，各国都不得不重视英美的战略性地位。

（2）翻译模型着重点不同：中国的研究机构将注意力的重点主要放在基于统计的机器翻译，特别是基于句法的统计机器翻译，对于基于实例的机器翻译研究相对较少；而日本对于基于实例的机器翻译研究比较深入。到底哪一种翻译模型更为优越，在不同的应用场景下应该有不同的选择。

（3）语言资源不同：机器翻译需要大量的语言学资源，包括双语语料库、各种术语词典或者标注树库。中国的各个研究机构在中文的资源建设方面已经积累了很多成果，日本则在日语的相关数据资源建设上更为成果卓著，这个事实是毋庸置疑的，因此如果能将两国的资源整合，或者两国合作开发更大规模的针对日语和汉语语言学特点的数据资源，将对于日汉和汉日的机器翻译发展起到极大的推动作用。

总之，日汉/汉日机器翻译与日英/英日机器翻译相比难度更大，汉语是一种属于孤立语的语言，与欧美语言以及日语相比，语法结构不同，没有活用的形与时态，句子不能以单词为单位进行切分，汉语句法结构分析等难度更大。但是，两国面向日汉和汉日的机器翻译研究已经存在，在CWMT2011的机器翻译评测中也已经设置了日汉的机器翻译项目。在此基础上进一步加强中日两国的面向日汉和汉日的机器翻译研究的合作对两国将是双赢的。

5 中日机器翻译合作研究展望

目前虽有一些汉日/日汉、英汉/汉英机器翻译系统，但这些系统的翻译准确率尚不如人意。因此，我们需要运用机器翻译方面最先进的研究成果、各种精确的词典、对齐的平行文本数据等制作优质的翻译系统。由于中国有中英/英中翻译系统，而日本有日英/英日翻译系统，因此也考虑以英语为桥梁，按照日—英—中、中—英—日的步骤来进行翻译。但是，即使日英、中英等系统的翻译准确率达到将近85%，如果以英语为桥梁进行翻译，则日中/中日的翻译准确率最高也只有70%。因此，为提高机器翻译的准确率，必须制作日文与中文的直接翻译系统。

中日两国在机器翻译领域展开合作具有较好的前期基础。JST与中国研究机构之间的深厚交流关系可谓由来已久。JST近年成立了中国综合研究中心，积极开展将中国的信息转为日语后进行传播的业务，并计划在今后积极地向中国传播日本的信息，并为此从2006年开始开发日汉/汉日机器翻译系统，投入9亿多日元并历时5年而于2011年3月初步完成系统开发，在翻译准确率方面日中翻译率约为90%，中日翻译率接近80%，与其他已有的系统相比毫不逊色。不过要将中日翻译系统推向实际运用，还需要进一步的努力与完善。

日本开发的系统以基于实例的机器翻译为框架，通过有效利用依存结构分析，吸收了语言表现的复杂变化，从而进行高质量的翻译。今后的工作主要是提高中文的依存结构分析精度，将实例词典扩充至现有系统词典数量的数倍，以期实现应对语言的多样性表现。为此，需要收集大量的对齐平行文本。此外，还要针对科技范畴内尽可能广泛的领域，完善其专业术语对译词典。这些工作单凭中国和日本研究人员的单方面努力终究有所局限。中国机器翻译主要在基于规则的机器翻译和基于统计的机器翻译方面进行研究，对中文句法分析等有深入研究。日中/中日机器翻译研究需要中日两国研究人员通过联合研究等形式展开合作，具体在对齐平行文本、实例词典、专业术语词典等共同课题展开研究。日方在过去5年间投入9亿多日元开发的有关中文的句法分析方法、各种词典、数据库对于各位来说也有可供利用的价值。

在目前基于规则、基于实例和基于统计的机器翻译系统进行不断融合之时，迫切需要中日两国在机器翻译领域加强合作，在日汉/汉日机器翻译技术与系统研制上进行优势互补，并首先在科学技术信息服务领域展开示范性应用，并逐步扩展到其他信息领域应用。

参考文献

[1] BROWN P, PIETRA S D, PIETRA V D, et al. The mathematics of statistical machine translation: parameter estimation [J]. Computational Linguistics, 1993, 19(2): 263-311.

[2] 长尾真.日中/中日机械翻译系统的发展[C]//中日两国机器翻译技术合作研讨会.北京,2011.
[3] 董振东.中国机器翻译的世纪回顾[N].中国计算机世界,2000年,第一期.
[4] 黄国文,张文浩.语文研究群言集[M]//周海中.机器翻译50年.广州:中山大学出版社,1997.
[5] 机器翻译的研究[OL]. [2011-10-13]. http://www.ia.ac.cn/kxcb/kpwz/200804/t20080414_2299218.html.
[6] NAGAO M. A framework of a mechanical translation between Japanese and English by analogy principle [C]// ELITHORN A, BANERJI R. Artificial and Human Intelligence. NATO Publications, 1984.
[7] FISCUS J G. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER) [C]// IEEE Workshop on Automatic Speech Recognition and Understanding, 1997.
[8] SCHWENK H, GAUVAIN J L. Improved ROVER using language model information [C]// ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millennium, Paris, Sep. 2000, 47-52.
[9] MATUSOV E, UEFFING N, NEY H. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment [C]// EACL, 2006.
[10] SIM K C, BYRNE W, GALES M, et al. Consensus Network Decoding For Statistical Machine Translation System [C]// ICASSP, 2007.
[11] ROSTI A I, ZHANG BING, MATSOUKAS S, et al. Improved Word-level System Combination for Machine Translation [C]// Proceedings of ACL 2007.
[12] CARPUAT M, WU DEKAI. Improving statistical machine translation using word sense disambiguation [C]// 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL 2007). Prague, Jun 2007: 61-72.
[13] CHAN YEE SENG, NG HWEE TOU, CHIANG D. Word sense disambiguation improves statistical machine translation [C]// 45th Annual Meeting of the Association for Computational Linguistics (ACL-07), Prague, Jun 2007.
[14] WU DEKAI, FUNG P. Can semantic role labeling improve SMT? [C]// 13th Annual Conference of the European Association for Machine Translation (EAMT 2009), Barcelona, 2009: 218-225.
[15] WU DEKAI, FUNG P. Semantic roles for SMT: A hybrid two-pass model [C]// Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009), Boulder, CO, 2009.
[16] FREDERKING R, NIRENBURG S. Three heads are better than one [C]// Proceedings of the fourth Conference on Applied Natural Language Processing, 1994: 95-100.
[17] WU DEKAI. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora [J]. Computational Linguistics, 1997, 23(3): 377-403.
[18] WU DEKAI, WONG H. Machine Translation with a Stochastic Grammatical Channel [C]// Proceedings of the 36th Conference of the Association for Computational Linguistics, 1998.
[19] XIONG DEYI, LIU QUN, LIN SHOUXUN. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation [C]// Proceedings of ACL 2006: 521-528.
[20] LIU YANG, LIU QUN, LIN SHOUXUN. Tree-to-String Alignment Template for Statistical Machine Translation[C]. In Proceedings of COLING/ACL 2006, Sydney, Australia, 2006:609-616.
[21] MI HAITAO, HUANG LIANG, LIU QUN. Forest-Based Translation [C]// Proceedings of ACL 2008, Columbus, Ohio, USA, 2008: 192-199.
[22] LIU YANG, LIU QUN. Joint Parsing and Translation [C]// Proceedings of COLING 2010, Beijing, China, 2010.
[23] 周玉,翟飞飞,张家俊,等.多语言文本翻译系统[C]//第七届全国机器翻译研讨会,厦门,中国,2011.
[24] 奚宁,赵迎功,汤光超,等.南京大学第七届机器翻译研讨会评测技术报告[C]//第七届全国机器翻译研讨会,厦门,中国,2011.
[25] 巢文涵,李舟军.ZZX_MT系统CWMT’2011评测报告[C]//第七届全国机器翻译研讨会,厦门,中国,2011.
[26] 何彦青,石崇德,于薇,等.中国科学技术信息研究所CWMT’2011技术报告[C]//第七届全国机器翻译研讨会,厦门,中国,2011.
[27] DEMPSTER A P, LAIRD N M, RUBIN D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm [J]. Journal of the Royal Statistical Society, Series B (Methodological), 1977, 39(1): 1-38.
[28] WANG YE-YI, WAIBEL A. Modeling with structures in statistical machine translation [C]// Proceedings of COLING-ACL’98, Montreal, Quebec, Canada, 1998: 1357-1363.
[29] OCH F J, TILLMANN C, NEY H. Improved Alignment Models for Statistical Machine Translation [C]// Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland, College Park, 1999: 20-28.
[30] OCH F J, NEY H. The alignment template approach to statistical machine translation [J]. Computational Linguistics, 2004(30): 417-449.
[31] KOEHN P, OCH F J, MARCU D. Statistical Phrase-based translation [C]// Proceedings of the Human Language Technology Conference and the North American Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, 2003: 127-133.
[32] KOEHN P, AXELROD A, MAYNE A B, et al. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation [C]// International Workshop on Spoken Language Translation, 2005.
[33] MARCU D, WONG W. 2002. A Phrase-based joint probability model for statistical machine translation [C]// Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002: 133-139.
[34] MARCU D, WANG WEI, ECHIHABI A, et al. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases [C]// Proceedings of EMNLP-2006, Sydney, Australia, 2006: 44-52.
[35] ZENS R, NEY H. A Comparative Study on Reordering Constraints in Statistical Machine Translation [C]// Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Sappora, Japan, 2003: 144-151.
[36] ZENS R, NEY H, WATANABE T, et al. Reordering Constraints for Phrase-Based Statistical Machine Translation [C]// Proceedings of the 20th International Conference on Computational Linguistics (COLING), Geneva, Switzerland, 2004: 205-211.
[37] OCH F J, THAYER I, MARCU D, et al. Arabic and Chinese MT at USC/ISI [C]// Presentation given at NIST Machine Translation Evaluation Workshop, 2004.
[38] TILLMANN C. A block orientation model for statistical machine translation [C]// HLT-NAACL, Boston, MA, USA, 2004.
[39] KUMAR S, BYRNE W. Local Phrase reordering models for statistical machine translation [C]// Proceedings of HLT-EMNLP, 2005.
[40] CHIANG D. A hierarchical Phrase-based model for statistical machine translation [C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, 2005: 263-270.
[41] GALLEY M, HOPKINS M, KNIGHT K, et al. 2004. What's in a translation rule? [C]// Proceedings of HLT-NAACL-2004, 2004: 273-280.
[42] ZOLLMANN A, VENUGOPAL A. Syntax Augmented Machine Translation via Chart Parsing [C]// NAACL 2006 - Workshop on statistical machine translation, New York, June 2006: 4-9.
[43] SHIEBER S M, SCHABES Y. Synchronous tree adjoining grammars [C]// Proceedings of the 13th International Conference on Computational Linguistics (COLING), 1990(3): 1-6.
[44] EISNER J. Learning non-isomorphic tree mappings for machine translation [C]// Proceedings of the 41st Annual Meeting of the Association for Computational

Linguistics, Poster/demonstration session, 2003: 205-208.
[45] MENEZES A, RICHARDSON S D. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora [C]// Proceedings of the Workshop on Data-driven Machine Translation, 2001: 39.
[46] MELAMED I D. Multitext grammars and synchronous parser [C]// Proceedings of HLT-NAACL, 2003: 79-86.
[47] MELAMED I D. Statistical Machine Translation by Parsing [C]// Proc. of ACL, 2004.
[48] FOX H J. Phrasal cohesion and statistical machine translation [C]// Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.
[49] LIN DEKANG. A path-based transfer model for machine translation [C]// Proceedings of the 20th International Conference on Computational Linguistics (COLING), Geneva, Switzerland, Aug 23-27, 2004: 625-630.
[50] DING YUAN, PALMER M. Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars [C]// Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL), Ann Arbor, 2005.
[51] QUIRK C, MENEZES A, CHERRY C. Dependency treelet translation: syntactically informed phrasal SMT [C]// Proceedings of proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL), Ann Arbor, Michigan, June, 2005: 271-279.
[52] SU KEH-YIH. To Have Linguistic Tree Structures in Statistical Machine Translation? [C]// Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE), Wuhan, China, October 30th-Novermber 1st, 2005.
[53] OCH F J, NEY H. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 2002: 295-302.

作者简介

张均胜，博士，计算机软件与理论专业，主要研究方向：多语言信息服务，语义计算。E-mail: zhangjs@istic.ac.cn
何彦青，博士，模式识别与智能系统专业，主要研究方向：机器翻译，自然语言处理。E-mail: heyq@istic.ac.cn
李颖，博士，信息系统专业。相关研究课题：知识组织、基于XML的数字出版、基于DOI的文献链接和DRM系统等，近期关注的话题为中国与日本“中日机器翻译”领域研发进展的比较研究——面向中日跨语言科技信息资源服务的国际合作。E-mail: liying@istic.ac.cn
王惠临，研究员，博士生导师，主要研究方向：多语言信息服务，机器翻译，自然语言处理。E-mail: wanghl@istic.ac.cn

Machine Translation Research in China and Japan: Advances and Comparison

Zhang Junsheng, He Yanqing, Li Ying, Wang Huilin / Information Technology Support Center, Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Machine translation investigates the use of computer software to translate text or speech from one natural language to another. Since the first computer was invented, people have been studying and exploring high quality and high efficiency of machine translation technology. Recently, rule-based machine translation, example-based machine translation and statistical translation are the main three translation patterns. There are some approaches of system combination for better machine translation results. With the development of science, technology, economy and culture, machine translation has become more important in breaking the language barrier between Chinese and Japanese for promoting China-Japanese exchanges and cooperation. Machine translation researchers in China and Japan have carried out a large number of Chinese-Japanese/Japanese-Chinese machine translation of theoretical research and system development. They have achieved a lot of effective results. however, it is still far from the practical translation application of large-scale and high quality. Therefore, it is necessary for researchers in China and Japan to cooperate in machine translation technology and system development for Chinese-to-Japanese and Japanese-to-Chinese, especially in the parallel corpus, dictionary, terminology, syntactic analysis and so on. This paper presents an overview of the China-Japanese machine translation research and compares machine translation research in China and Japan. We also discuss the prospects of China-Japanese cooperation in machine translation research.

Keywords: Machine translation, Rule-based machine translation, Example-based machine translation, Statistical machine translation, System combination

（收稿日期：2011-11-03）