特约专题 SPECIAL ARTICLES
doi:10.3772/j.issn.2095-915x.2017.03.002

扫描二维码与作者深度对话

面向科技文献的中日机器翻译合作研究

中国科学技术信息研究所 北京 100038
赵志耘 石崇德 何彦青 高影繁 姚长青

摘 要 本文以多语言科技信息服务为立足点，结合中日两国面向科技文献的机器翻译研究现状，介绍了两国近几年开展的机器翻译合作项目的情况，包括合作背景与基础、知识产权、具体合作内容与成果，以及在机器翻译实用化方面的一些思考。

关键词： 机器翻译，科技文献，国际合作
中图分类号： G35, TP391

Cooperative Research on Chinese-Japanese Machine Translation for S&T Documents

Institute of Scientific and Technical Information of China, Beijing 100038, China
ZHAO ZhiYun SHI ChongDe HE YanQing GAO YingFan YAO ChangQing

Abstract Based on the practical use in multi-lingual S&T information service, this paper introduced the temporary situation in machine translation research in S&T document between China and Japan. The paper mainly presented the cooperative research project on machine translation, which involved the background, cooperative foundations, copyright problem and the achievements. The paper also gave some ideas about how to promote practical use of machine translation in S&T information service.

Keywords: Machine translation, S&T document, international cooperation

基金项目： 本文受国家自然科学基金青年项目（71403257），国家创新方法工作专项（2015IM020500）的资助。
作者简介： 赵志耘（1966–），女，博士，研究员，研究方向：战略与决策管理；石崇德（1979–），通讯作者，博士，副研究员，研究方向：自然语言处理、机器翻译，Email：shicd@istic.ac.cn；何彦青（1974–），博士，副研究员，研究方向：机器翻译，自然语言处理，Email：heyq@istic.ac.cn；高影繁（1974–），博士，副研究员，研究方向：多语言信息处理，Email：gaoyingf@istic.ac.cn，知识组织；姚长青（1974–），副研究员，博士，研究方向：情报理论与方法，Email：yaocq@istic.ac.cn。

1 引言

随着全球化的飞速发展，不同国家不同语种的信息交流越来越频繁，全球的翻译市场规模稳步增长。据统计，2014 年全球翻译行业市场规模达到 371.9 亿美元，比上年同期增长 6.23%，预计到 2020 年全球翻译产业规模将达到 530 亿美元左右 [1]。越来越多的科技公司开始通过机器翻译来介入这个庞大的市场。

机器翻译（Machine Translation）是使用计算机把一种语言翻译成另外一种语言的一门科学 [2]，以解决不同语言之间的交流障碍。随着近两年神经网络机器翻译兴起，机器翻译的质量得到显著的提升，谷歌、百度、有道、腾讯、搜狗等众多科技公司都开展了机器翻译产品的研发，一些实时机器翻译系统开始在体育、旅游等领域得到应用。

外文科技文献是科研人员和科技信息工作者获取国际科研动态的重要来源，由于其包含大量专业领域词汇，信息获取难度较大。日本是一个科技强国，在科技领域有很多值得学习的地方，但是由于我国科研人员很少具备日语语言能力，导致对日本的科技发展不能进行深入了解和充分借鉴。以国家科技图书文献中心的文献服务为例，到 2015 年为止共收录日文文献共 400 多万篇，涵盖了日本国内出版的大部分重要科技文献，实际 2015 年国内原文传递量不足 3%，远低于英文文献的服务数量。在这种情况下，利用机器翻译能够辅助科研人员从外文文献中挖掘重要信息，进而掌控国际科研动态。

近十多年随着我国综合国力的快速提升，日本也越来越重视与我国在多领域的交流与合作，其中语言问题是影响双方交流的主要障碍。中日科技机器翻译合作研究正是在这一背景下开展。中日两国的机器翻译研究都有比较长的历史，在各自语种的分析和处理方面有丰富的经验，合作研究能更好的取长补短，减少单方面的研发投入，加快研发进程。机器翻译合作研究是中日两国开展科技领域合作的一项重要课题，其最终目标是推动两国机器翻译研究共同进步，同时实现在科技文献翻译方面的实用化。

2 中日科技机器翻译研究现状

我国是继美国、前苏联、英国之后，世界上第四个开展机器翻译研究工作的国家。1956 年，机器翻译研究就列入了我国科学工作的发展规划，1959 年进行了俄汉机器翻译实验 [3]。目前国内在机器翻译研究方面已经达到世界领先水平。

相对于通用领域的机器翻译研究，科技领域的机器翻译研究相对较少。中国科学技术信息研究所（简称中信所）很早就参与了机器翻译研究工作，主要面向机器翻译在科技信息工作中的应用。中信所于 1974 年参加“748”工程（汉字信息处理系统工程），启动机器翻译研究；1978 年 12 月与社会科学院语言研究所合作研发的“英汉题录机器翻译”投入试用；1987 年参与了日本 ODA 多语言机器翻译研究；1989 年参与“智能型英汉机器翻译系统（IMT/EC-863）项目”并获得国家科技进步一等奖；十一五期间，中信所开展了机器翻译在科技文

doi:10.3772/j.issn.2095-915x.2017.03.002

献信息处理方面的应用研究，在全国机器翻译研讨会的科技类机器翻译评测中名列前茅，英汉科技机器翻译系统已经服务于国家科技图书文献中心网站的英文论文标题和摘要翻译[4-6]。

日本在机器翻译领域的研究也非常早，同时一直致力于推动亚洲语言的机器翻译研究[7,8]，并且提出了基于实例的机器翻译理论[9]，推动了机器翻译理论方法的进步。2006 年，为了促进科技信息的传播，日本开始实施“日中·中日语言处理技术开发研究”项目，探索运用语言结构的通用型基于实例的翻译方法，研究从对译语料中半自动生成科技文献翻译和信息检索用辞书的方法，特别以中文为焦点，构筑大规模语言资源，利用该语言资源开发机器翻译原型系统、进行实验及评价[10]。与中国不同的是，日本的企业研究所占研究机构的 60% 以上，东芝、Retec 等公司也投入大量资源开展机器翻译研究。尤其随着专利数据的快速增长，专利领域的机器翻译也同样是日本机器翻译研究的重点，由日本国立情报学研究所组织的 NTCIR 评测中，专利机器翻译是重要的评测任务之一[11]。

3 合作研究背景

在科研全球化背景下，中日机器翻译的开发研究对促进中日科技信息的双向传播、促进中日科技战略合作、促进中日创新的可持续发展至关重要[12]。日本从 2006 到 2010 年的 5 年间，政府累计投资科研经费近 10 亿日元，通过相关研究所和大学的合作，实施了国家重要项目“日中·中日语言处理技术开发研究”，但是仍然有部分问题无法得到很好的解决，包括日汉机器翻译双语资源不足、中文句法分析准确率较低、缺少科技领域机器翻译实际应用等。

为了解决这些问题，日本科学技术振兴机构 JST 开始在中国寻找日汉机器翻译研究合作伙伴，并于 2011 年 6 月向当时的科技部曹健林副部长进行了汇报，表达了合作意愿。科技部对本项合作非常支持，通过中信所联络国内在中文信息处理和机器翻译领域具有丰富经验的科研单位与日方进行了多次项目研讨，共同协商中日两国开展面向科技文献的机器翻译合作研究的各种问题，尤其是需要解决的核心问题、技术路线、知识产权以及如何开展实用化等问题。经过多方共同努力，科技部国际合作专项“面向科技文献的日汉双向实用型机器翻译合作研究”于 2014 年 2 月正式获批，执行时间为 2014 年 4 月至 2016 年 3 月，中方由中信所牵头，联合中科院自动化所、哈尔滨工业大学及北京交通大学共同参与，日方由 JST 牵头，联合京都大学、筑波大学、丰桥技术科技大学及情报通信研究机构等单位共同参与。

4 合作研究基础与知识产权

日本是传统的科技强国，近几年连续多年获得诺贝尔奖，其科研实力可见一斑，其科技动态和科研经验值得我国科研工作者关注和借鉴；随着近些年我国综合科研能力的提高，日本国内科研部门也非常关注我国的科研动向，希望从中获取有价值的信息。中信所与 JST 作为中日两国重要的科技信息机构，各自都需要为本国科研人员和决策部门提供国外科技动态，

这一共同目标是双方开展科技领域机器翻译合作项目的基础。

中日两国机器翻译研发人员分别在各自语言的句法分析技术上具有天然优势。日方在基于实例的机器翻译引擎、日语词法和句法分析方面处于国际领先位置，近年来中方在统计机器翻译方面取得了长足的进步，参与本项目合作的中科院自动化所在 IWSLT、NIST 机器翻译评测中名列前茅，哈尔滨工业大学在中文句法分析研究方面也有很深的积淀。双方合作能够取长补短，充分利用对方已有的技术降低己方的投入，同时提升自己的服务水平。

由于合作项目涉及中日两国多家单位，代表着不同的利益团体，为了保证合作的顺利进行，双方各成员单位首先对合作项目中可能产生的的知识产权问题进行了界定。中日双方本着平等互换的原则在合作研究框架内共享研究成果，主要内容涉及以下四点：(1) 理论研究成果公开发表；(2) 合作发表的学术论文按贡献大小排列作者及其单位；(3) 独立完成的专利、软件著作权和学术论文属各自国家的单位；(4) 在项目合作范围内，项目参与单位可以自由使用其他单位提供给项目的数据和工具，合作项目结束之后，则需要征求原单位同意才可继续使用。

总之，共同的目标、互补的技术以及清晰的知识产权界定，是本合作项目能够顺利开展的最基本条件。

5 合作研究内容与成果

本合作项目主要目标是建立实用型日汉双向机器翻译系统并在科技信息服务领域进行应用，合作内容包括双语资源建设、日汉词法句法分析工具及机器翻译引擎的共建和共享。

机器翻译研究是一项系统工程，涉及双语资源、模型方法和工具、具体应用等三个层次，图 1 展示了整个中日机器翻译合作项目的研究路线图。

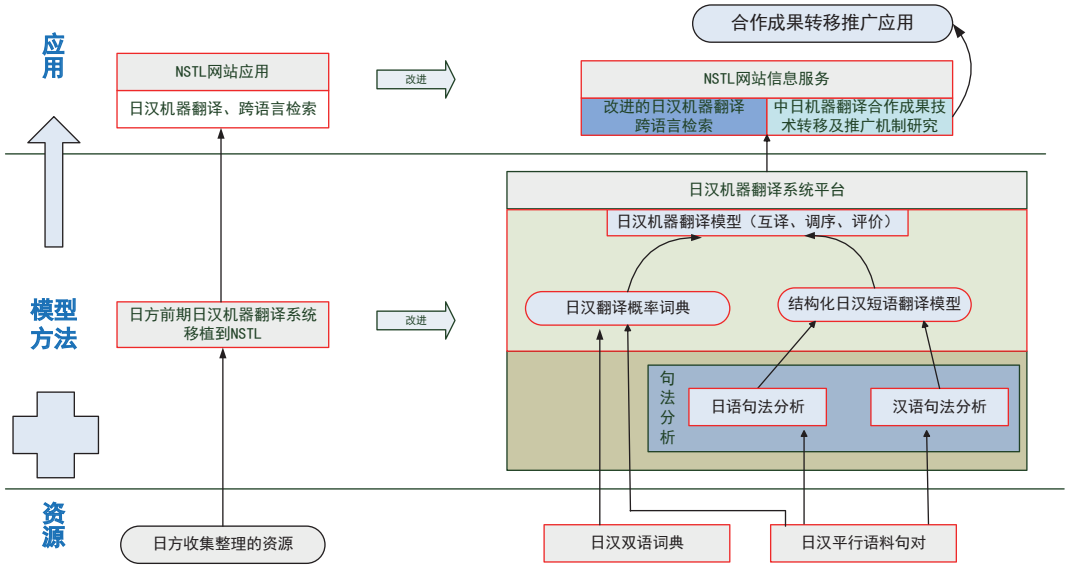

图1 中日机器翻译合作研究路线图

doi:10.3772/j.issn.2095-915x.2017.03.002

其中语言资源是整个项目的基础，包含日汉双语词典和日汉双语平行句对。平行句对是统计机器翻译训练必须的资源，规模越大，翻译引擎的翻译效果越好。由于科技文献中专业词汇较多，不同科技领域的文献内容差别较大，这对科技领域机器翻译双语资源提出了更高的要求，因此本项目通过科技翻译词典来对双语语料进行一定程度的补充。同时，双方通过共建共享的方式进行合作，能够大大降低单独一方的投入。

模型方法部分是合作研究的重点，涵盖了机器翻译系统的各个模块，包括中文和日文的词法分析、句法分析、翻译引擎研究等。具体执行过程中，中日双方分别由中信所和 JST 负责统筹，对每家合作单位的具体工作内容进行了分工，并定期对工作进度进行监督。除了邮件交流之外，每年均安排两国研究人员出国互访，进行面对面交流，一方面能更好的理解彼此的研究工作，共同商议项目和难点的解决办法，另一方面也增进了项目成员之间的友谊。经过合作研究，日方基于实例的机器翻译引擎在项目结束时在汉日翻译中达到世界领先水平；中方的科技领域中文句法分析研究也采用了最先进的分词、词性标注及句法分析的一体化模型，句法分析准确率由 80% 提升到 90%；中方在翻译引擎研究方面，分别在如何充分利用源语言和目标语言句法知识方面、如何有效引入实例机器翻译思想方面以及如何利用语义信息方面提出了多种创新思想，结合深度学习在机器翻译引擎方面的应用，大大提高了机器翻译引擎的翻译效果。

成果应用方面，合作项目最终由中信所和 JST 分别在科技信息服务中开展集成应用示范。中信所将项目成果应用于国家科技图书文献中心（NSTL）的日语论文题目和摘要翻译服务，JST 则将汉日机器翻译应用于日本最大的科技文献信息服务网站 JDREAM 所收录的中文论文摘要的翻译。如图 2、3 是分别是 NSTL 和 JDREAM 网站应用的示例。

图2 NSTL日语论文摘要翻译

图3 JDREAM中文论文摘要翻译

6 结论

机器翻译这样的跨语言研究适合通过国际合作来进行，尤其是面向科技的机器翻译更偏

向公共服务的性质，两国合作一方面可以减少单个国家在资源建设方面的投入，另一方面可以充分利用双方在本国语言处理方面的专长，取长补短，加快实现在科技信息服务中的实用化，增进两国的科技和文化交流。

中日两国的机器翻译合作项目一开始就从合作目标、合作模式、合作内容、知识产权等多方面进行了较为详细的设计和协商，为顺利开展合作研究奠定了良好的基础。项目过程中双方一直都秉承平等互利的原则，双方研究人员在合作过程中建立起深厚的友谊。项目的合作模式、合作成果及合作过程中遇到的问题和解决方案，为开展更多跨国机器翻译合作研究提供了宝贵的经验。

目前国际合作项目已经顺利达成了预期目标，并完成结题验收。但是对参与课题的中日两国研究机构和研究人员来说，为期两年的合作项目实际上只是双方开展机器翻译合作研究的一个开始，中信所与 JST 签署了后续的合作协议，继续推进双方在科技领域机器翻译实用化方面的工作，JST 和日本国内的科研机构也更多的认识到国际合作的优势，因此在近几年不断的通过 NTCIR、WAT 等会议来推动研究机构进行亚洲语言机器翻译的研究。近两年随着神经网络和深度学习方法在机器翻译中的应用，机器翻译的效果越来越好，中信所与 JST 不仅致力于日汉机器翻译在公益类科技信息服务中的应用，同时也开始在商业应用方面进行更深入的合作，继续推进日汉机器翻译的实用化。

语言障碍一直以来都是制约不同国家和民族进行文化交流的主要障碍，每年全球有几百亿美元用于翻译，机器翻译研究正不断努力来解决这个问题。相信在不久的将来，机器翻译会更多的走入人们的生活，并帮助科研人员实现有效无障碍的跨语言科技交流。

参考文献

[1] 博思数据 . 全球翻译产业 [EB/OL]. [2017-2-14]. http://www.bosidata.com/qitawenjiaoshichang1601/493271R2Q7.html.

[2] 冯志伟 . 机器翻译研究 [M]. 北京 : 中国对外翻译出版公司 , 2004.

[3] 冯志伟 . 机器翻译与语言研究 [J]. 术语标准与信息技术 , 2007(4): 38-41.

[4] 何彦青 , 石崇德 , 于薇 , 张均胜 , 王惠临 . 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论文集 . 厦门 , 2011: 81-87.

[5] 何彦青 , 石崇德 , 张均胜 , 王惠临 . 中国科学技术信息研究所 CWMT'2013 技术报告 [C]. 第九届全国机器翻译研讨会论文集 . 昆明 , 2013.

[6] 张均胜 , 何彦青 , 李颖 , 王惠临 . 中日两国机器翻译研究进展及比较 [J]. 数字图书馆论坛 , 2011(12): 20-31.

[7] 冯志伟 . 机器翻译 —— 从实验室走向市场 [J]. 语言文字应用 . 1997(3): 73-78.

[8] Nakazawa T, Ding C, et al. Overview of the 3rd Workshop on Asian Translation[C]// Proceedings of the 3rd Workshop on Asian Translation (WAT2016). Osaka, Japan. 2016: 1-46.

[9] Nagao M. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle[J]. Artificial and Human Intelligence, 1984: 173-180.

[10] 李颖 , 吴琳 . 从跨语言信息检索论日本国日中 · 中日机器翻译研发前沿 [J]. 数字图书馆论坛 , 2008(9): 42-47.

[11] Patent Machine Translation Task at NTCIR-9[EB/OL]. [2017-2-14]. http://ntcir.nii.ac.jp/PatentMT/.

[12] 乔晓东 . 中日两国机器翻译技术合作研讨会 [J]. 数字图书馆论坛 , 2011(12): 1-2.