郭肖肖, 李子林, 刘庆猛, 等. 基于文献计量的机器翻译研究综述[J]. 智能计算机与应用,2025,15(9):107-111. DOI:10.20169/j.issn.2095-2163.250917

基于文献计量的机器翻译研究综述

郭肖肖, 李子林, 刘庆猛, 李雪山
(中国铁道科学研究院集团有限公司 科学技术信息研究所, 北京 100081)

摘 要: 为系统反映国内机器翻译领域研究现状、热点和前沿,为研究人员了解机器翻译领域研究提供概览性参考,本文应用可视化软件 Citespace 对 2013~2022 十年间中国知网数据库中有关机器翻译的文献进行文献计量和知识图谱分析。研究发现:国内关于机器翻译领域的研究呈波动上升趋势,计算机界、语言学界分别从机器翻译技术、翻译教学、译文质量、机器翻译应用场景等角度开展了广泛研究。从研究热点来看,主要集中在统计机器翻译、神经机器翻译、跨语言检索、译后编辑的深度研究上。其中,针对资源稀缺领域的机器翻译问题,神经网络、注意力机制、迁移学习和领域适应是学者们下一步研究的重点。

关键词: 机器翻译; 文献计量; 知识图谱
中图分类号: TP391.2 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0107-05

A review of machine translation research based on bibliometrics

GUO Xiaoxiao, LI Zilin, LIU Qingmeng, LI Xueshan
(Scientific and Technological Information Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: To systematically reflect the current status, hotspots, and frontiers of machine translation research , and to provide an overview reference for researchers to understand the field, this paper utilizes Citespace for a bibliometric and knowledge mapping analysis on machine translation from the CNKI database spanning 2013~2022. Research finds that domestic research on machine translation shows a fluctuating upward trend. The computer and linguistic communities have conducted extensive research from the perspectives of machine translation technology, translation teaching, translation quality, and machine translation application scenarios. From the perspective of research hotspots, the main focus is in-depth research in statistical machine translation, neural machine translation, cross-language retrieval, and post-editing. Among these, for machine translation in resource-scarce areas, neural networks, attention mechanisms, transfer learning, and domain adaptation are the key focuses of scholars' future research.

Key words: machine translation; bibliometrics; knowledge graph

0 引 言

机器翻译是指使用计算机将文本从一种自然语言(源语言)转译成另一种自然语言(目标语言),是计算语言学领域的一个分支,融合了语言学、翻译学、数学和计算机科学等多学科在内的应用技术和研究方法。从基于规则的方法到基于语料库的策略(包括基于实例和基于统计的方法),再到基于神经网络的技术,机器翻译已经经历了多个研究阶段。伴随着人工智能技术的进步、计算机处理能力的增强和多语言信息量的迅速增加,机器翻译的研究领域正在不断拓展和深化,展现出巨大潜力和可观前景。

1 研究方法及数据来源

1.1 研究方法

文献计量学是以文献作为研究对象,通过应用数学和统计学方法来分析和评估科学研究的现状及

基金项目: 中国铁道科学研究院集团有限公司基金课题(2021YJ133)。
作者简介: 李子林(1995—),女,博士,高级工程师,主要研究方向: 信息资源管理; 刘庆猛(1992—),男,助理研究员,主要研究方向:计算机技术研发; 李雪山(1976—),男,研究员,主要研究方向:计算机技术研发,信息资源管理。
通信作者: 郭肖肖(1991—),女,助理研究员,主要研究方向:信息资源管理,信息分析。Email:1316170515@qq.com。
收稿日期: 2023-12-22

其发展动向。运用文献计量学对特定学科领域的论文进行趋势分析和内容深度挖掘,可以客观地绘制出该领域的研究现状和发展趋势图景,为研究人员和科研决策者规划研究方向和制定科研策略提供参考依据。

Citespace 是一款信息可视化工具,专门用于分析和可视化共引网络,从而帮助用户洞察特定领域的前沿知识和发展趋势,并进行文献的科学计量分析。本文选用的分析工具为 CitespaceVI,利用其可视化功能,对收集到的数据进行处理,绘制了近十年来国内机器翻译领域研究的知识图谱,梳理国内机器翻译研究年度发文量、作者、机构等信息,挖掘研究热点,展示国内机器翻译领域的研究趋势和动向。

1.2 数据来源

本文以 2013~2022 年中国知网(CNKI)数据库中的期刊论文为数据基础,以“机器翻译”或“计算机翻译”为主题词进行检索,检索得到期刊论文共计 2 414 篇,通过对检索到的文献进行人工筛选,剔除非学术性的新闻报道、会议通知、广告等与主题不相关或不密切的文献,最终得到中文文献 2 222 篇。

2 文献基本情况分析与讨论

2.1 发文量分析

文献发文数量的时间序列分布是衡量某一领域研究发展趋势的重要指标,发文量随时间分布图可直观反映出不同时期该领域的研究热度随时间变化情况。2013~2022 年国内机器翻译研究相关文献的年度分布情况及变化趋势如图 1 所示。

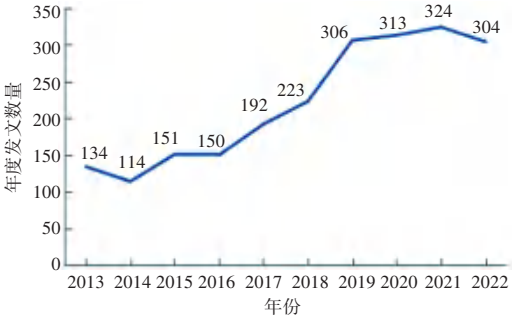

图 1 2013~2022 年国内机器翻译研究相关文献的年度分布情况及变化趋势

Fig. 1 Annual distribution and trend of literature related to machine translation research in China from 2013 to 2022

由图 1 可知,国内机器翻译领域的研究呈现出波动性增长的趋势[1]。近年来,学术界对机器翻译的关注和研究兴趣在持续上升。具体来说,可分为 3 个阶段:波动发展阶段(2013~2016 年)、快速发展阶段(2016~2019 年)和稳步发展阶段(2019~2022 年)。在 2013~2016 年,机器翻译领域的发文量相对稳定。2016 年以后,随着人工智能技术的迅速发展和机器翻译技术的优化,发文量显著上升。特别是在 2016 年,谷歌发布了神经机器翻译系统 GNMT,运用人工神经网络提高翻译质量,为机器翻译技术带来了重大突破。同年,微软的 Microsoft Translator 开始采用神经网络进行语音翻译。2017 年,网易推出了以 NMT 系统为核心的有道神经网络翻译系统 YNMT,专注于中文与其他语言的互译。2018 年,科大讯飞发布了讯飞翻译机 2.0,与此同时,微软开发的机器翻译系统在中译英的新闻报道测试中达到了与人类专业译者相当的水平。此外,搜狗翻译、阿里巴巴翻译和腾讯翻译也在这一时期迅速发展。自 2019 年以来,机器翻译领域的年发文量保持在每年约 300 篇的水平,但与之前相比增长速度有所放缓,这可能与该时期机器翻译技术的成熟发展有关。

2.2 主要发文机构

机构是学科研究领域的间接主体,机构发文量统计及占比可直观反映该机构在领域内研究地位,发文量前十的研究机构见表 1。

表 1 国内机器翻译领域研究发文机构 Top 10

Table 1 Top 10 posting institutions in the field of machine translation in China

序号	机构	发文量
1	苏州大学计算机科学与技术学院	53
2	昆明理工大学信息工程与自动化学院	41
3	中国科学院大学	41
4	新疆大学信息科学与工程学院	34
5	中国科学院新疆理化技术研究所	24
6	昆明理工大学云南省人工智能重点实验室	23
7	内蒙古工业大学信息工程学院	21
8	中国科学技术信息研究所	21
9	广东外语外贸大学	13
10	上海外国语大学	11

由表 1 可以看出,发文机构多为理工类院校的计算机学院、信息学院,以及信息技术、人工智能实验室,其次是外语类院校。在研究方向上,理工类院校在机器翻译领域研究多侧重于机器翻译技术研究。其中,新疆大学、内蒙古工业大学在少数民族语言机器翻译研究领域特色鲜明;外语类院校在机器翻译领域研究多侧重于翻译教学、译文质量、机器翻译技术应用场景等。

2.3　主要研究者

作者是学科研究领域的直接主体，作者数量统计反映该领域研究热度；作者发文统计及占比可反映该作者在领域内的重要研究地位。通过分析，共计 410 位国内学者发表了与机器翻译相关的文献，发文量前十的作者见表 2。

表 2　国内机器翻译领域高产作者 Top 10

Table 2　Top 10 productive authors in the field of machine translation in China

序号	作者	发文量
1	余正涛	40
2	杨雅婷	27
3	苏依拉	19
4	艾山·吾买尔	18
5	仁庆道尔吉	18
6	王华树	18
7	段湘煜	17
8	熊德意	17
9	李军辉	16
10	李晓	14

由表 2 可以看出，国内机器翻译领域高产作者列表中大多为来自计算机领域的学者，代表作者包括余正涛、杨雅婷等，余正涛、杨雅婷、苏依拉、艾山·吾买尔、仁庆道尔吉在低资源语言翻译、汉越、汉泰、维汉、蒙汉机器翻译领域开展了深入研究。其次为语言学界学者，代表作者有王华树，主要研究方向为译后编辑、翻译教育、译者素养等。机器翻译为综合性、跨学科研究领域，涉及语言学、翻译学、数学、计算机科学等多个学科。计算机界学者在机器翻译技术研究方面具有优势，语言学界在翻译教学、译文质量、应用场景方面具有优势，二者应加强学科间互动合作，进而提升机器翻译领域研究的深度，推动机器翻译的多领域应用普及和优化。

2.4　研究热点

将中文文献数据导入 Citespace 中，得到国内机器翻译研究的关键词共现网络图。就关键词出现次数而言，频次 top15 为：机器翻译（757）、人工智能（147）、神经机器翻译（122）、计算机辅助翻译（103）、自然语言处理（91）、译后编辑（90）、翻译技术（86）、统计机器翻译（68）、深度学习（64）、人工翻译（62）、神经网络（57）、语料库（53）、注意力机制（45）、翻译教学（35）、大数据（28）。在关键词网络图谱基础上聚类结果如图 2 所示，形成了 13 个聚类，依次为：#0 深度学习、#1 机器翻译、#2 翻译技术、#3 计算机辅助翻译、#4 统计机器翻译、#5 循环神经网络、#6 机器学习、#7 平行语料库、#8bert、#9gru、#11 跨语言检索、#12 跨语言信息检索、#14 自我认同。通过关键词和聚类分析，可以得到机器翻译领域研究内容主要集中在 3 个方面，分别是：机器翻译技术研究、机器翻译场景应用、机器翻译质量评价。

图 2　国内机器翻译领域关键词聚类

Fig. 2　Keyword clustering in the field of machine translation in China

（1）机器翻译技术研究。包含聚类标签深度学习、翻译技术、计算机辅助翻译、统计机器翻译、循环神经网络、机器学习、平行语料库、bert、gru。这些聚类主要集中在统计机器翻译和神经机器翻译两大技术领域。

① 统计机器翻译。统计机器翻译通过对大量

的平行语料进行统计分析,构建统计翻译模型进行翻译,有效减少了对人工的依赖,Google 翻译的大部分语言对采用的是统计机器翻译的方法。由于统计机器翻译依赖于巨大的语料库,在应对数据稀疏、平行语料较少时显得不足;此外,模型中所包含的句法、语义成分较低,在处理句法差别较大的语言对时存在问题,如何提高统计机器翻译的领域自适应能力已然成为学界的研究热点。

② 神经机器翻译。鉴于统计机器翻译在数据稀疏、复杂语言现象及翻译规则方面的不足,神经网络机器翻译逐渐成为当前机器翻译系统的主流方法。神经机器翻译通过使用深度学习神经网络获取自然语言之间的映射关系,可以从大规模平行语料中自动学习翻译规则和语言表示,直接将源语言句子映射到目标语言句子。此外,其还可根据之前数据输入与分析结果进行数据的自动升级与优化,不断提升计算能力。为解决神经机器翻译在处理长句子和复杂句子结构上的困境,学者们引入了注意力机制。注意力机制允许模型在翻译过程中关注源语言句子中不同位置的信息,对源语言的不同部分进行加权关注,使模型在翻译时集中关注源语言句子中与当前正在翻译的目标语言位置相关的信息,更好地理解源语言句子的重要部分[2],从而提高翻译质量。

③ 平行语料库。统计机器翻译和神经机器翻译都需要以语料库为基础,平行语料库的收集和构建直接影响机器翻译的效果。随着翻译形式和内容的多样化,在应对资源稀缺语言对,如蒙汉、维汉、藏汉等,以及新领域资源翻译的问题日益凸显,零资源翻译逐渐成为机器翻译领域一个研究方向。零资源翻译是指在没有源语言和目标语言之间的平行语料的情况下进行翻译,学者们研究通过数据增强[3-4]、多语言的预训练模型、自监督学习[5]、迁移学习[6]和领域适应[7]等方法,解决零资源翻译的挑战[8],提高模型在不同领域的翻译性能。

(2)机器翻译场景应用。包含聚类标签跨语言信息检索、跨语言检索。

① 跨语言检索。随着走出去、“一带一路”战略实施,国际交流日益繁盛,跨语言信息检索的重要性日益凸显。机器翻译技术上通过将查询翻译成目标语言、来实现跨语言的信息检索,实现了不同国家语言之间的智能翻译。此外,研究人员还探索了将知识图谱与跨语言信息检索相结合的方法来提高跨语言检索的准确性和丰富性[9]。在场景应用方面,学者们探索了跨语言信息检索在搜索引擎、少数民族地区政务服务[10]、智慧图书馆的研究及应用[11],基于机器翻译的语音识别[12-13]、机器翻译机器人[14]、机器翻译系统也成为机器翻译产品化的研究方向。

② 多模态翻译。多媒体数据的快速增长催生了机器翻译领域多模态翻译研究的发展,多模态翻译旨在将图像、视频、音频等多模态输入与文本翻译相结合,实现跨模态的翻译任务。研究人员致力于图像与文本之间的对齐、融合图像视觉注意力机制[15]、融合以及利用视觉信息来改善翻译质量,从而实现不同应用的,包括语音识别翻译、图像翻译、视频翻译、手语翻译等。

(3)机器翻译质量评价。包含聚类标签译后编辑。

① 译后编辑。机器翻译缺少人工翻译对句子逻辑、语句构造等进行判断,对于创造性高的文本,如文学、诗歌,以及专业性强的领域等,其译文质量有待提升。学者们探索了翻译错误自动检测、人机交互等译后编辑技术来提高机器翻译的准确性。译后编辑技术包括自动译后编辑 APE、人工译后编辑[16],自动译后编辑本质上与机器翻译相同,由于学习了纠错规则,其翻译质量比机器翻译要高,相较于人工译后编辑效率更高。人工翻译、机器翻译与译后翻译效果相比,译后翻译与机器翻译在结果上更为相似,在原文有一定难度且机器翻译译文质量不高的情况下,译后编辑译文质量不如人工翻译[17]。

② 机器翻译质量评估研究。学者们探索了多种算法对机器翻译质量进行评估,如 bleu、TER、METEOR、APE 等。其中,bleu、TER 需要参考人工译文,在没有译文的情况下如何开展机器翻译质量评估,经历了基于特征工程和机器学习、基于深度学习、融合预训练语言模型方法三个研究阶段[18],学者们提出了一系列不需要参考人工译文、由机器自动对机器翻译译文进行评估的方法、如 APE。

2.5 研究前沿

研究热点随时间变化趋势可以看出不同时间研究热点的变化,可以为把握研究热点、分析研究趋势提供基础。本文绘制了基于时间序列的国内机器翻译研究趋势演进图,如图 3 所示。由图 3 可以看出该领域研究热点总体上从宏观的围绕机器翻译、计算机辅助翻译、译后编辑、语料库、翻译技术处理,逐步向深度学习、强化学习、预训练、注意力机制、多模态翻译、稀缺语言翻译等细分领域研究拓展。

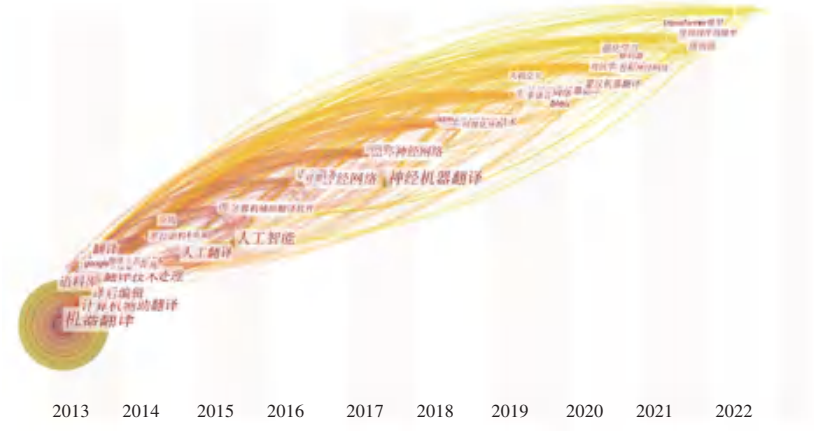

图 3　国内机器翻译领域研究时序图

Fig. 3　Timing diagram of research in the field of machine translation in China

分析图 3 可知，2013～2015 年，学者们的研究内容广泛，包括机器翻译、计算机辅助翻译、译后编辑、语料库、翻译技术处理，这些方向大致与之后的研究方向相同，在后续研究中不断地对各体系范畴进行深入探索。2016～2018 年，随着人工智能的发展，尤其是神经网络技术的应用，机器翻译领域研究不断深化，逐渐衍生出深度学习等研究方向。2019～2022 年，强化学习、预训练、注意力机制、多模态翻译[19]、稀缺语言翻译逐渐成为学者们的热点研究方向[20]。

3　结束语

本文借助可视化软件 Citespace 对 2013～2022 十年间中国知网数据库中有关机器翻译的文献进行了文献计量和知识图谱分析，从发文量、发文机构、发文作者、研究热点及前沿等方面揭示了当前国内机器翻译领域研究现状。

本文主要基于中文文献对机器翻译领域研究现状进行评述，未来可对国外机器翻译领域研究进行分析，对比国内外研究热点、研究趋势，为研究人员更全面把握机器翻译领域研究热点和前沿提供基础。

参考文献

[1] 穆军芳，张丽鑫. 国际机器翻译近十年的动态演进：基于 CiteSpace 和 VOSviewer 的可视化分析[J]. 沈阳大学学报(社会科学版)，2022，24(6)：643-654.

[2] 石磊，王毅，成颖等. 自然语言处理中的注意力机制研究综述[J]. 数据分析与知识发现，2020，4(5)：1-14.

[3] 侯宏旭，孙硕，乌尼尔. 蒙汉神经机器翻译研究综述[J]. 计算机科学，2022，49(1)：31-40.

[4] 朱俊国，杨福岸，余正涛，等. 低频词表示增强的低资源神经机器翻译[J]. 中文信息学报，2022，36(6)：44-51.

[5] 袁扬，李晓，杨雅婷. 基于 LDA 主题模型的维吾尔语无监督词义消歧[J]. 厦门大学学报(自然科学版)，2020，59(2)：198-205.

[6] 李洪政，冯冲，黄河燕. 稀缺资源语言神经网络机器翻译研究综述[J]. 自动化学报，2021，47(6)：1217-1231.

[7] 刘欢，刘俊鹏，黄锴宇，等. 面向低资源俄汉机器翻译的领域适应方法[J]. 厦门大学学报(自然科学版)，2022，61(4)：654-659.

[8] 张文博，张新路，杨雅婷，等. 面向低资源神经机器翻译的回译方法[J]. 厦门大学学报(自然科学版)，2021，60(4)：675-679.

[9] 昆明理工大学. 融合领域知识图谱的汉越跨境民族文本检索方法及装置[P]. CN：202211350058，2023-01-13.

[10] 赵生辉，陈刚. 民族地区政务大厅多语言服务环境构建策略研究[J]. 价值工程，2018，37(17)：279-283.

[11] 刘莉，王怡，邵波. 面向智慧图书馆的多语言自动翻译平台架构设计研究[J]. 图书馆学研究，2022(6)：37-44.

[12] 郭慧骏. 基于人工智能技术和语音识别的机器同步翻译系统[J]. 现代电子技术，2022，45(9)：152-156.

[13] 李俊. 计算机翻译辅助技术在同传中的应用及对同传生态系统的影响[J]. 中国翻译，2020，41(4)：127-132.

[14] 叶楠，寇丽杰. 多语言机器人深度学习模型构建[J]. 信息与控制，2020，49(6)：680-687.

[15] 李霞，马骏腾，覃世豪. 融合图像注意力的多模态机器翻译模型[J]. 中文信息学报，2020，34(7)：68-78.

[16] 孟福永，唐旭日. 效率为先：机器翻译译后编辑技术综述[J]. 计算机工程与应用，2020，56(22)：25-32.

[17] 张威，明昊. 人工翻译、机器翻译与译后编辑的对比实证分析：以汉语介词结构翻译为例[J]. 沈阳师范大学学报(社会科学版)，2021，45(5)：114-120.

[18] 邓涵铖，熊德意. 机器翻译译文质量估计综述[J]. 中文信息学报，2022，36(11)：20-37.

[19] 吴友政，李浩然，姚霆，等. 多模态信息处理前沿综述：应用、融合和预训练[J]. 中文信息学报，2022，36(5)：1-20.

[20] 林倩，刘庆，苏劲松，等. 神经网络机器翻译研究热点与前沿趋势分析[J]. 中文信息学报，2019，33(11)：1-14.