Modern Linguistics 现代语言学, 2024, 12(3), 287-292
Published Online March 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.123202

Hans汉斯

国内近二十年机器翻译错误研究综述

刁丽昱，蔡良钰，仰玉静

中国矿业大学(北京)文法学院，北京

收稿日期：2024年2月7日；录用日期：2024年3月20日；发布日期：2024年3月29日

摘　要

本文以CNKI (中国知网)期刊数据库中收录的关于机器翻译错误研究的论文为数据来源，运用数据统计的研究方法，对国内机器翻译错误研究进行可视化分析，通过考察论文发表数量、主要研究领域、机器翻译引擎、错误标注手段以及错误分类方法等，整理分析了国内机器翻译错误的发展过程及现状。研究结果表明，国内机器翻译错误研究在最近十年兴起并呈上升趋势，在2021年达到高潮，文本聚焦领域比较广泛。文章进而提出了神经网络机器翻译错误技术研究存在的问题，展望未来研究导向，以促进国内机器翻译错误类型再研究、再创造。

关键词

机器翻译，研究热点，译文错误

A Review of Machine Translation Error Research in China in the Past Two Decades

Liyu Diao, Liangyu Cai, Yujing Yang

School of Law and Humanities, China University of Mining and Technology (Beijing), Beijing

Received: Feb. 7th, 2024; accepted: Mar. 20th, 2024; published: Mar. 29th, 2024

Abstract

This paper takes the papers on machine translation error research included in the CNKI database as the data source, and applies the research method of data statistics to visualize and analyze the domestic machine translation error research, and organizes and analyzes the development process and the current situation of the domestic machine translation error by examining the number of papers published, the main research fields, the machine translation engine, the means of error annotation, and the method of error classification, and so on. The research results show that domestic machine translation error research has emerged and shown an upward trend in the last

文章引用: 刁丽昱, 蔡良钰, 仰玉静. 国内近二十年机器翻译错误研究综述[J]. 现代语言学, 2024, 12(3): 287-292.
DOI: 10.12677/ml.2024.123202

decade, reaching a climax in 2021, with a relatively wide range of text-focused fields. The article then puts forward the problems existing in the research of neural network machine translation error technology and looks forward to the future research orientation, in order to promote the domestic machine translation error type re-study and re-creation.

Keywords

Machine Translation, Research Hotspots, Translation Errors

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

1949 年 Weaver 与 Shannon 在合作出版的《通讯的数学原理》一书中，将思想的交流看作“编码—解码”过程[1]。Weaver 首次提出了用计算机辅助翻译的方法，为机器翻译理论的发展奠定了坚实的基础，此后机器翻译研究得以兴起并快速发展。根据陈善伟[2]的研究，其发展大致可以分为以下三个阶段：1967~1983，初步发展期；1984~1992，稳步发展期；1993~200，快速发展期。机器翻译经过几十年的发展，已经逐渐得到人们的认可和信任。国家出台《新一代人工智能发展规划》，将人工智能上升至国家战略；世界正在经历“百年未有之大变局”，国与国之间频繁持续的交流互动驱使语言服务需求激增；《2022 中国翻译及语言服务行业发展报告》显示，语言服务企业普遍认同“机器翻译 + 译后编辑”的工作模式，机器翻译在行业的应用越来越广泛，机器翻译的重要性和必要性日益彰显[3]。国内对于机器翻译错误研究最早开始于 1985 年，近十年来，机器翻译错误研究呈逐年上升趋势，论文发表数量在 2021 年达到顶峰。对于中英机器翻译错误类型的研究，研究文本较为广泛，但数量上还略显不足。目前，国内中英机器翻译错误类型的研究文本包含科技领域，政务领域，外宣文本，专利文献领域，网络文学作品领域，石油化工领域。尽管机器翻译技术取得了长足发展，机器翻译质量得到了大幅提升，但是与人工翻译质量相比，机器译文质量尚存在较大差距，机器翻译还面临着许多挑战与问题。

2. 研究方法及数据来源

本研究采用了文献研究法和描述性研究法，以中国知网数据库为数据来源，以“机器翻译错误”为检索主题，共得到文献 325 篇，且国内有关机器的翻译失误的研究最早出现在 1985 年。考虑到数据的可靠性和代表性，经筛选和去重得到的有效文献总数为 276 篇，时间跨度为 1998~2022 年。

3. 数据分析与讨论

3.1. 研究文献概况

以 CNKI 论文库为基础，研究者获取了年度论文数量如图 1 所示。

总体而言，国内关于机器翻译失误的研究处在不断上升阶段。高璐璐、赵雯[1]提出国内的翻译机器错误研究的发展可以划分为两个阶段：缓慢发展时期(1998~2016)，这一时期的论文数量比较少，由于前期机器翻译的应用投资及技术研究的大范围停滞，这一时期的研究还处于恢复阶段。但这一时期机器翻译领域也发表了比较重要的一些文献：例如，2004 年，刘春燕发表《论科技文体的翻译原则与方法》，

对科技文本的翻译原则与方法进行分析概括；2012 年，罗季美，李梅发表《机器翻译译文错误分析》，主要对机器译文和人工译文进行对比研究，分析机器翻译的典型错误；2013 年，李梅，朱锡明《英汉机译错误分类及数据统计分析》研究机译典型错误并进行比较完整数据统计等。相对快速发展时期(2016~2022)，由于大量机器翻译平台和多语语料的产生和发展，机器翻译得到更加广泛的运用，与此同时，机器翻译所产生的错误翻译也得到更多的关注，所以这一时期的论文数量逐渐增多。

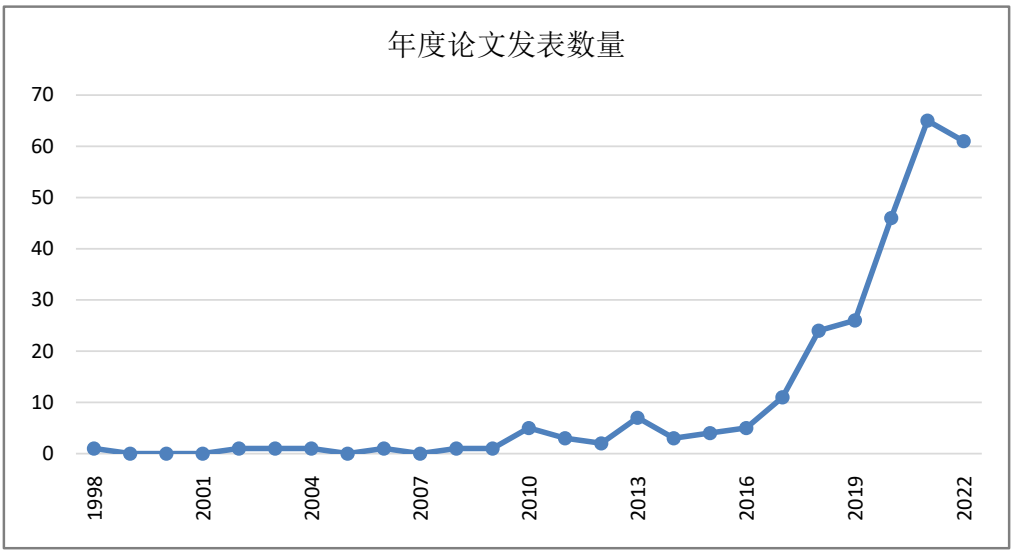

Figure 1. Diagram of annual number of papers published on machine translation error research
图 1. 机器翻译错误研究年度论文发表数量题图

3.2. 研究的文本领域

经过数据整理发现，机器翻译错误研究的学科领域变得越来越广泛(图 2)。

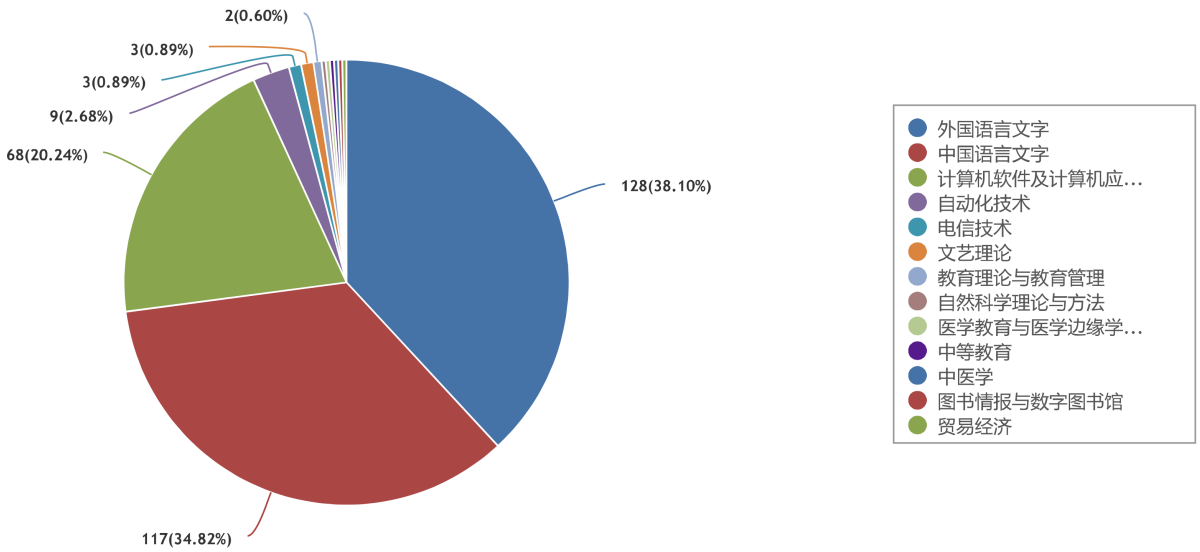

Figure 2. Diagram of fields involved
图 2. 涉及领域题图

从图 2 可以得出，机器翻译失误领域研究主要聚焦于经济领域，例如：北京外国语大学张珏于 2020 年发表《浅析英汉机器翻译错误类型及其译后编辑策略》等；自然科学领域：于艳英，聂蜂林于 2019 年发表《基于科技文本的在线机器翻译错误分析》等；文学领域：刘霄驰，杜宇娇于 2022 年发表《网络文学作品汉译英机器翻译的错误类型研究》等；医学领域：唐斌，陈烁于 2020 年发表《在线机器翻译软件的医学文本翻译评析》等；政治领域：蔡欣洁，文炳于 2021 年发表《汉译英机器翻译错误类型统计分析——以外宣文本汉译英为例》、上海外国语大学罗佳琦于 2022 年发表《探讨政务类文本机器翻译的错误类型及解决对策》等；以及能源领域内的石油文本机器翻译失误。关于机器翻译文本的选择，国内研究主要集中在信息类文本以及科普类文本等。

信息类文本主要起到传递信息和反映客观事实的作用，它的语言一般不带个人色彩，信息来源可靠，专业性较强，对于研究机器翻译失误的研究空间较大。科普类文本旨在传递科学知识，所以这类文本在翻译的过程中需要大量的专业和文化知识积累。同时，该类文本中含有大量的专业术语以及长难句，这些都是在翻译过程中容易出现错误的地方。所以科普类文本对于研究机器翻译错误具有较大的可选择性。

3.3. 机器翻译引擎

对中国知网输入“机器翻译错误”关键词检索出的 327 篇文章中的核心期刊进行分类整理后，又对其中机器翻译引擎进行了整理统计，笔者发现研究者使用的机器引擎种类丰富，其中谷歌翻译的使用量最高，机器翻译引擎的使用也体现了研究者的喜好，例如孙逸群[4]进行的材料类文摘机器辅助翻译错误剖析和海洋类论文摘要的机器翻译辅助翻译错误剖析[5]都使用了谷歌作为翻译引擎，李奉栖[6]在进行人工智能时代人机英汉翻译质量对比研究也使用了谷歌作为检测人工译员是否抄袭机器译本的机器翻译引擎。在前几年中谷歌翻译比较准确，研究者们大都使用其作为翻译引擎，但自从 2022 年谷歌翻译退出中国市场之后就鲜有研究者以其作为翻译引擎。研究者们还使用 DeepL，腾讯翻译，金山 AI 等翻译引擎，其中，罗季美、李梅[7]在两次研究汽车文本机器翻译错误时均使用的中国科学院计算机语言信息工程研究中心研发的华建机器翻译系统以及百度、有道、词霸、谷歌、必应等翻译引擎，这些翻译引擎的主要特点是市场占有率比较高、使用者比较广泛，其中，百度翻译是国内市场份额第一的翻译类产品，广泛受到研究者们的青睐。除此之外，近些年来，人工智能得到了迅猛发展，之后的研究者可以关注使用 ChatGPT 等新兴翻译引擎来进行机器翻译错误的相关研究。

3.4. 机器译文错误类型及标注手段

机器翻译译文以及人工翻译译文之间的平行对比是一个工作量很大的工程，因此有些研究者会借助一些机器自动标注译文错误平台的帮助，但是自动标注的准确性不高，因而有些研究者会采取自动标注错误和人工标注错误相结合的方法，例如裘白莲，王明文，李茂西等[8]在进行构建“细粒度英汉机器翻译错误分析语料库”就采用了自动标注和人工标注相结合的方式。不过，绝大多数研究者为保证机器翻译错误标注的准确性还是采用了人工标注的方法，比如罗季美、李梅[7]在进行机器翻译译文错误分析、在进行机器翻译句法错误分析以及李奉栖[6]在进行人工智能时代人机英汉翻译质量对比研究都使用了人工标注的方法。

同时为了进一步了解机器翻译错误的发展状况，笔者统计了相关核心期刊中所出现的翻译错误的主要类型。虽然不同学者研究的方向有所不同，但是目前关于机器翻译错误的研究主要集中在术语，准确度，语言规范，本地化规范等方面。比如，孙逸群[4] [5]侧重于研究术语类翻译错误；裘白莲，王明文等[8]侧重于对准确度的研究，主要包括词语增译，漏译，错译以及语序的错误；而罗季美[9]在《机器翻译句法分析》中主要侧重于对语言规范的机器错误分析。同时，罗季美，李梅[7]以及还研究了符号的错误

翻译，符号机器翻译错误也需注意[10]，但相较于其它类型的错误而言，这一错误类型的研究目前还比较少。

同时，词汇是组成句子的基本元素，词汇翻译对译文的质量有着至关重要的影响，词汇错译发生率所占比例远远高出其他类型的译文错误。罗季美,李梅[7]也指出众多研究都重点分析机器翻译在词语方面的错误。

4. 问题与总结

由上述以“机器翻译错误”为关键词进行的核心文献统计分析结果表明，我国在机器翻译错误类型领域的研究取得了不错的成果。研究的文本领域不断扩大，呈现出多元化的特征；机器翻译引擎的选用范围扩大，不拘泥于热点翻译平台如百度翻译、有道翻译和谷歌翻译，而是呈现“海纳百川，不局限一家之言”的局面，在广泛分析不同多样平台的基础上寻找机器翻译的错误类型共性；错误类型研究的不断细化和有序分类，从词语，句法，文本和符号等层面进行分析和总结；在研究方法上也更加精细，虽然机器标注相较于人工更加快捷便利，但是国内学者更多偏向于选择机器标注和人工标注相结合，或仅采用人工标注，体现了科学严谨的学术风格。然而，尽管国内学者在机器翻译错误类型领域进行了不断地探索，成果颇丰，然而，与美国、英国等最早开展相关研究的国家相比，我国还存在着许多不足，值得反思[11]。

其一，文本研究领域具有局限性。近十年来，国内在机器翻译错误类型方面开展了较为广泛的研究，如语言文字、计算机、自动化技术、文艺理论、电信技术、教育、自然科学、航空航天科学与工程、医学、贸易经济等领域，对机器翻译在各类领域中可能出现的错误类型进行了较为深刻的分析与探讨，但是在能源矿业领域翻译失误研究方面仍然有较大的研究空缺。我国物产丰富，是能源矿业资源的富藏大国，也是能源消费大国，机器翻译研究者们还需要加强这一领域的研究。

其二，研究的机器翻译引擎类型具有局限性。对机器翻译领域的研究大多集中在各类机器翻译平台如谷歌翻译平台上。然而，需要真正了解和分析机器翻译容易出现的错误类型，完善机器翻译语料库并进一步提升机器翻译能力，不能局限于一个或少数几个翻译平台的问题研究。通过对近十年机器翻译错误领域论文的研究，发现几乎大多数论文的研究对象主要及集中在谷歌翻译，必应翻译，搜狗翻译，有道翻译，百度翻译，金山翻译，腾讯翻译等在国内外较为流行的七个翻译平台上，有部分研究也只集中在百度，谷歌等翻译平台上。近年来，随着人工智能和大数据产业的不断崛起，ChatGPT 一举兴起，成为 AI 行业的爆款产品。它的出现影响了众多行业的发展航向，其中，ChatGPT 更是在翻译领域大展身手，成为翻译行业的关注点。机器翻译错误类型研究动向应紧随时代发展，紧跟时代要求。

其三，研究理论与技术缺乏深度结合。通过分析发现，目前国内翻译技术研究并不均衡。一是真正探讨翻译技术的学者相对较少。研究机器翻译、译后编辑、本地化翻译等的还是固定的人群[12]。大多数国内学者对于机器翻译错误类型研究仅仅停留在理论层面，真正深入到技术层面进行机器翻译错误类型研究的学者少之又少。一是因为多数掌握理论的学者不了解技术，多数掌握技术的人无法深入了解理论知识；刁洪[11]指出语言学者不懂计算机，计算机专家又不精通外语。“道”与“器”咫尺天涯，人文与技术难以共融。此外，学界与业界存在隔阂，少见学术机构与商业公司的科研合作。

基金项目

本文受中国矿业大学(北京)大学生创新训练项目(202308021)资助。

参考文献

[1] 高璐璐, 赵雯. 机器翻译研究综述[J]. 中国外语, 2020, 17(6): 97-103.

[2] 陈善伟. 翻译科技新视野[M]. 北京: 清华大学出版社, 2014(6): 68-73.

[3] 戴光荣, 刘思圻. 神经网络机器翻译: 进展与挑战[J]. 外语教学, 2023, 44(1): 82-89.

[4] 孙逸群. 材料类文摘机助翻译的错误剖析[J]. 中国科技翻译, 2018, 31(3): 15-18.

[5] 孙逸群. 海洋类论文摘要机辅翻译错误剖析[J]. 中国科技翻译, 2019, 32(2): 31-33+30.

[6] 李奉栖. 人工智能时代人机英汉翻译质量对比研究[J]. 外语界, 2022(4): 72-79.

[7] 罗季美, 李梅. 机器翻译译文错误分析[J]. 中国翻译, 2012, 33(5): 84-89.

[8] 裘白莲, 王明文, 李茂西, 等. “细粒度英汉机器翻译错误分析语料库”的构建与思考[J]. 中文信息学报, 2022, 36(1): 47-55.

[9] 罗季美. 机器翻译句法错误分析[J]. 同济大学学报(社会科学版), 2014, 25(1): 111-118+124.

[10] 夏玲, 李宜蔓, 李弘武. 人工智能背景下科技论文摘要的机器翻译与译后编辑[J]. 编辑学报, 2022, 34(4): 396-401+406.

[11] 刁洪. 国内翻译技术研究综述[J]. 北京第二外国语学院学报, 2017, 39(6): 69-81+125.

[12] 唐悦妮, 梁满玲. 基于 12 种外语类核心期刊的可视化分析研究——国内翻译技术研究现状综述(1980-2021) [J]. 现代信息科技, 2021, 5(20): 122-126.