Modern Linguistics 现代语言学, 2022, 10(9), 1944-1949
Published Online September 2022 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2022.109261

Hans汉斯

机器翻译的译文质量、高频错误类型及解决对策研究：基于机器翻译的发展史

韦佑武[1]，李　娜[1*]，赵良威[2]

[1]东北电力大学，吉林 吉林
[2]中国能源建设集团天津电力建设有限公司，天津

收稿日期：2022年8月15日；录用日期：2022年9月14日；发布日期：2022年9月21日

摘　要

智能信息时代下机器翻译在语言服务行业已经成为提升译者效率与译文质量不可或缺的工具。本文在概括机器翻译的发展历史、主要发展阶段中机器翻译出现的翻译质量问题及相关的技术瓶颈的基础上，重点关注机器翻译高频出现的错误类型，继而归纳相应的解决方案，以期为科技文本机器翻译译后编辑的高频错误或偏误类型提供一个相对较为全面的分析与系统研究。

关键词

机器翻译，错误类型，解决对策

Translation Quality, High-Frequency Error Types and Solutions of Machine Translation: Based on the Development History of Machine Translation

Youwu Wei[1], Na Li[1*], Liangwei Zhao[2]

[1]Northeast Electric Power University, Jilin Jilin
[2]China Energy Engineering Group Tianjin Electric Power Construction Co., Ltd., Tianjin

Received: Aug. 15th, 2022; accepted: Sep. 14th, 2022; published: Sep. 21st, 2022

[*]通讯作者。

文章引用: 韦佑武, 李娜, 赵良威. 机器翻译的译文质量、高频错误类型及解决对策研究: 基于机器翻译的发展史[J]. 现代语言学, 2022, 10(9): 1944-1949. DOI: 10.12677/ml.2022.109261

Abstract

Machine translation has become an indispensable tool for improving translators' efficiency and translation quality in the language service industry in the era of intelligent information. Based on the development history of machine translation, translation quality problems and related technical bottlenecks in machine translation in the main development stages, this paper focuses on the types of errors that occur frequently in machine translation, and then summarizes the corresponding solutions, with a view to providing a relatively comprehensive analysis and systematic study of the types of high-frequency errors or biases in the post-translation editing of scientific and technical texts by machine translation.

Keywords

Machine Translation, Error Types, Solutions

1. 引言

全球化进程和信息化技术快速发展的背景下，海量信息资源不断推出和迅速更新给翻译工作带来了前所未有的挑战，机器翻译(Machine Translation，简称 MT)迎合了云计算、大数据、深网络的发展趋势，在语言服务行业已经成为提升译者效率与译文质量不可或缺的辅译工具，改变了译者的传统翻译工作方式，逐步发展成为应对该挑战的重要方式和手段。

2. 机器翻译的工作原理和适用性

机器翻译的工作原理是模仿人工翻译所遵循的过程模式，其步骤可概况为原文输入、原文分析、译文综合和译文输出。但是，翻译是涉及人类系统化思维的一项复杂的动态活动，翻译的具体过程更是多源多层次的互动。首先，接收到的源语言信息需通过译者的推理与判断，实现源语深层结构的概念与目标语深层概念之间的匹配，再根据目标语的表达规则重新转换、组合与对接目标语表层信息的具体结构及形式，最高程度地保持源语言与目标语在语义与语用方面的一致性。

学界一般认为，机器翻译适用于程式化的信息类文本，具有用词固定、词汇语义信息投入精确、语句结构重复率较高的典型特征[1]，通常以描述事实、传递信息、知识和观点等文本为主[2]。

3. 机器翻译的发展阶段及相应的瓶颈技术问题

3.1. 机器翻译的发展阶段

机器翻译的思想发轫于 20 世纪 30 年代初法国科学家阿尔楚尼(G.B. Artsouni)，20 世纪 40 年代自然语言处理之父 Weaver 首次提出机器翻译的概念，即“用解读密码方法指导机器翻译”[3]。

机器翻译诞生以来经过七十余年的发展，其发展主要经历了四个阶段，即：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、基于深度学习的神经机器翻译。基于规则的机器翻译(Rule-based MT)，即基于词典和语法生成翻译，在 20 世纪 50 年代到 80 年代实现翻译机器化的早期发展

阶段，并提出了机器翻译最基本的工作原理及运行模式；基于实例的机器翻译(Example-based MT)技术的诞生及应用是基于互联网技术发展之上的产物，主要依赖于平行文本或多语料提供的大量可靠翻译实例，但基于实例的机器翻译技术发展受限于数据及低资源语言实例的短缺或稀少问题；基于统计(Statistics-based MT)的机器翻译，需通过大规模大量的平行语料构建统计翻译模型，进而使用此模型进行翻译，因此受限语料数量与分类精度，无法适用于典型的低资源语言；神经机器翻译(Neural machine translation, NMT)是基于海量信息数据和深度学习的技术，可通过建立多语言词汇共享词表进行语义映射共享语义空间，扩充了语言模型参数，解决了由于语言差异大而引发的词表共享受限问题，可实现在共享的同时又不丢失语言本身的多样性，共享资源利用人工神经网络直接将源语句映射为译句，缩小语言之间的语义距离，即以端到端的方式进行翻译建模。我国最早于 1957 年对机器翻译的研究正式开始，2013 年以来 NMT 快速发展，加拿大蒙特利尔大学机器学习实验室发布了神经网络的机译系统 GroundHog，2015 年百度发布了将统计和自动学习相结合的在线机译，2016 年机器翻译走入人工智能技术与机器翻译融合的阶段，国内外出现了如谷歌翻译、百度翻译、微软必应、网易有道、DeepL、SDL Trados Studio、阿里翻译等典型神经网络机器翻译。

机器翻译的发展从早期基于词和句法信息、基于实例的机器翻译，过渡到大规模的实例统计模型翻译和正在融合深度学习(Deep Learning)的神经机器翻译，每个阶段的发展趋势都见证了机器翻译的精确性与译文整体质量的逐步提升的同时，也显露了其阶段发展的瓶颈问题。

3.2. 机器翻译的发展瓶颈

机器翻译的发展过程是寻求认知、计算和技术之间的最高程度的拟合。拟合的程度高低决定了机器翻译的质量。早在 1966 年，自动语言处理咨询委员会(Automatic Language Processing Advisory Committee, ALPAC)调查报告表明机器翻译的译文质量低于人工翻译。时至今日，机器翻译的发展仍存在诸多瓶颈。

首先，语料库建设的规模和滞后问题。语料库是机器翻译赖以发展的驱动力之一，因此其规模、领域范畴影响着机器翻译的效率和质量。然而，语料库的建设本身也面临各种难以解决的问题。规模问题。只有语料库规模达一定规模的词条时，神经网络翻译的性能才开始优于统计翻译。语料库低于此规模时，机器翻译的优势无法凸显。目前语料库也是主要集中在时政新闻和科学技术等方面，绝大多数其他领域的语料库都严重缺乏；滞后性问题。科技领域中，每天都有大量的新生专业术语产生。此外，日常生活中新的表达方式也不断地被创造出来。这些新生术语或表达方式的译文产生需要一个过程，无法被语料库收集。

其次，机器翻译的增强技术问题。第一、以句子为输入单位，即所谓“端到端”的翻译方法，简单来讲就是指机器学习时以句子为单位进行输入，在输出端同样得到以句子为单位的译文。但是如果句子偏长的话，机器翻译就难以理清其中的逻辑关系。目前，机器翻译技术仍然停留在句法阶段，未能实现技术的完全语义化[4]；第二、语义与语法偏差，由于翻译以及语言本身的复杂性，导致不同语言在翻译过程中产生语义及语法模糊不清。语言是文化的载体，在机器翻译过程中很难准确把握其内在含义，继而更难将语义精确表述。同样，语法结构差异化也会导致在机器翻译过程出现语法错误。

最后，缺乏跨语言情感域能兼顾语音、语义深层信息和情感特征信息的重构能力。虽然深度学习技术可以处理深层次的语言信息并实现数据自动存储和技术升级，但机器翻译只能够学习人类的部分逻辑思维能力，且无法学习包括情感和想象力在内的形象思维。由于神经机器翻译的训练数据的难度，以及不能完全依靠机器的自学能力，深度学习技术支持下的神经机器翻译也面临着许多挑战[5]。

4. 机器翻译的高频错误类型和解决策略

根据 2017 年 ISO 18587 质量标准认证，机器翻译被定义为使用计算机程序将一个自然语言的文本自

动翻译到另一个自然语言，而译后编辑为“对机器翻译输出结果的编辑和修改”[6]。对比目前机器翻译和人工翻译的质量，虽然某些特定领域的文件的机器翻译基本达到了人工翻译的水平，但仍存在很多问题，所以需要通过扩大数据库和译后编辑等手段加以修改和补充，即增加机器翻译结果的编辑和修正过程，使通过译后编辑后的译文达到预期标准，进而提高整体翻译质量。

4.1. 机器翻译的高频错误类型

对机器翻译的高频错误类型的分析和研究脉络，以国内主流翻译研究类期刊《上海翻译》、《中国翻译》、《中国科技翻译》在 CNKI 数据库的数据为检索平台，以“机器翻译错误”为检索式选择文献进行机器高频错误类型述评。其中，基于统计学的机器翻译问题主要包括由于语言文化差异导致的双语口吻不一致问题[7]，以及过度翻译、欠译、形式错误、格式错误、多译漏译、冗余、词性判断错误、从句翻译错误以及短语顺序错误等[8]；基于 Google 翻译等神经网络翻译平台，由于正在发展的神经网络翻译技术可以在多个机器翻译的结果中自动给出较高质量的译文，只需要修正机器输出结果中对读者来说不易阅读和理解的部分，该部分包含语言实际使用时的主被动形式的选择、同反义词的替换、短语的搭配、时态的一致、缩略语的使用以及各类文本格式的调整等，可以代替译者解决大部分重复性翻译工作。Google 科技文本汉英机器翻译的错误主要出现在词汇、句法、逻辑等方面，且出现错误频次递减；错误发生频次如图 1 所示。

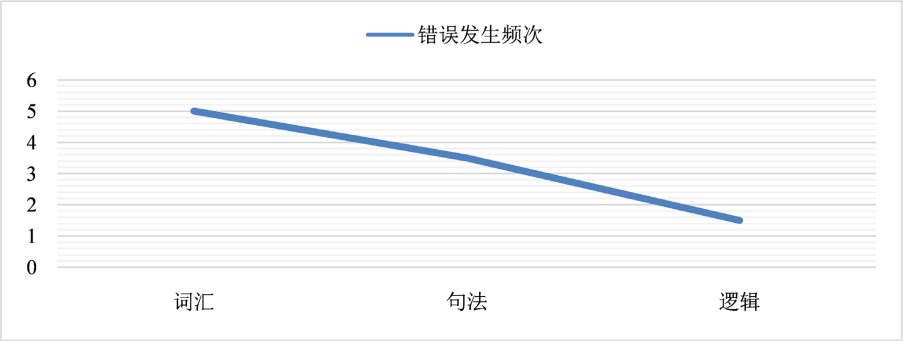

Figure 1. Frequency of errors
图 1. 错误发生频次

对比腾讯翻译和 Google 翻译平台的机器翻译效果显示出神经网络机器翻译存在较多术语错误、语义错误、单复数及冠词错误、前后不一致、漏译、主语缺失等问题[9]，并表明在所有错误类型中，术语错误、语义错误在所有错误类型中占比最高[10]，具体错误类型分布如图 2 所示。

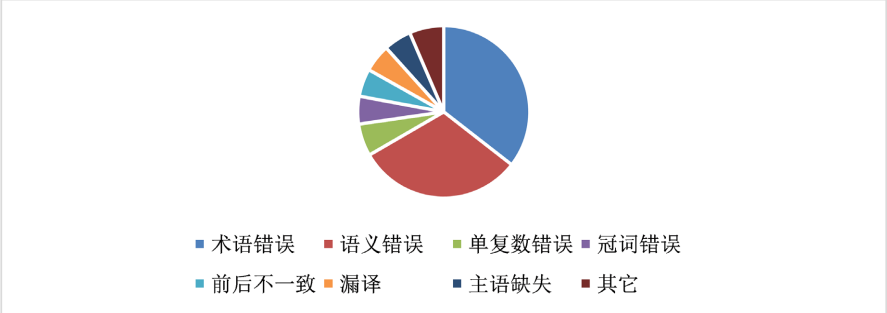

Figure 2. Distribution Chart of error types
图 2. 错误类型分布图

部分研究者采用的二级错误分类体系[11],并参考“中国高校外语专业多语种语料库建设与研究英语语料库”的错误分类标准[12],把错误类型分为两个大的层面:词语层面主要包括:词义偏差、中文语境导致的词义偏差、冗余、出现完全重复的英文、介词逻辑错误等;句段层面主要包括:被动语态使用错误以及过于紧贴原文结构造成的中式表达。

综上所述,笔者把错误类型主要分为四个方面:第一、词语层面,主要包括术语错误、词义偏差、欠译、误译、漏译等;第二、语法层面,主要包括短语搭配、句子结构、主语缺失、介词误用、主动/被动语态等;第三、句段层面,主要包括语序混乱、句法不一致、逻辑混乱等;第四、其它错误类型,包括标点符号、图表、格式误用等。以上主要错误类型如表 1 所示。

Table 1. Main error types
表 1. 主要错误类型

词语	语法	句段	其它
术语错误	短语搭配	语序混乱	标点
词义偏差	句子结构	句式糅杂	数字
欠译	主语缺失	语义错误	图表
误译	介词误用	句法不一致	符号
漏译	主动/被动语态	逻辑混乱	格式

4.2. 高频错误类型的主要解决方案

机器翻译的主流研究倾向于在剖析现阶段机器翻译底层机制的基础上,对机器翻译的发展方向提出设想,同时对译后编辑的方法、策略提出新要求。由于汉英语言差异以及机器翻译技术的局限性,机器翻译的错误类型复杂多样,可总体划分为词语层面和句段层面。针对上述所提及的错误类型,相应的建议对策如下:

第一、建立大规模的数据库,并设计具有针对性的专用语料库。尤其是在缺少足够语料的文本类型,如文学类文本、哲学类文本、低资源语种文本,一般机器翻译软件在翻译过程中主要是将文章分析与句子重组输入到系统当中,难以对词语的含义等进行细致分析,无法消除歧义。解决语义误译或错译问题,应设计针对性的专用语料库,收集词语与词语之间的固定搭配,将语言特征与更加结构化的知识存储连接起来,增加邻接语义的条件依赖和连续语义增强的相关数据,从而逐渐消除翻译歧义问题。

第二、增强文本数据的识别和标注技术。在建立基础数据库的同时,引入汉语知识增强文本数据,需要吸收来的其他感官数据和知识来增强,根据标注样本生成更多的标注样本以实现集成的多模态数据标注,提高语言差异性的识别率,克服隐喻等隐性特征所导致的弱语义表达现象。

第三、提升机器翻译译后编辑(Post-editing 或 PE)的策略。机器翻译结合译后编辑是目前保证译文质量的重要手段,可以在保证翻译效率的同时提升译文质量。机器翻译缺乏形象思维能力以及翻译过程出现语义语法偏差等,因此需结合译后编辑来纠正错误和调整句子结构,使译文更加符合译入语的语言环境。

5. 结语

从最初基于规则的机器翻译到现今的深度学习神经网络机器翻译,机器翻译历经多个发展时期,机器翻译的效率和质量迎合了日益增长的翻译需求。本文简述了机器翻译的发展历史、发展瓶颈、错误类型及其解决对策四个方面,机器翻译在技术方面仍在不断取得突破和完善,对于译者而言,使用机器翻

译系统所提供的各种功能对译者综合能力变得越来越重要。

基金项目

本论文部分研究获得全国教育科学教育部青年项目(EIA160479)和吉林省高教研究课题(JGJX2021D116)的支持。

参考文献

[1] Hutchins, W. and Somers, H. (1992) An Introduction to Machine Translation. Academic Press, London.

[2] 胡开宝, 李翼. 机器翻译特征及其与人工翻译关系的研究[J]. 中国翻译, 2016, 37(5): 10-14.

[3] Poibeau, T. (2017) Machine Translation. The MIT Press, Boston.

[4] Shannon, C.E. (1956) A Universal Turing Machine with Two Internal States. In Shannon, C.E. and McCarthy, J., Eds., *Automata Studies*, Princeton University Press, New Jersey, 157-165.

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

[6] (2014) ISO TC 37. ISO 18587 Translation Services—Post-Editing of Machine Translation Output—Requirements.

[7] 刘晓燕. 科技翻译中的“口吻”传递[J]. 中国科技翻译, 2014, 27(1): 12-15.

[8] 崔启亮, 李闻. 译后编辑错误类型研究——基于科技文本英汉机器翻译[J]. 中国科技翻译, 2015, 28(4): 19-22.

[9] 蔡强, 董冬冬. 基于 GOOGLE 神经网络汉英翻译的译后编辑研究——以科技文本为例[J]. 西南石油大学学报(社会科学版), 2020, 22(1): 107-112.

[10] 杨玉婉. 神经机器翻译的译后编辑——以《潜艇水动力学》英汉互译为例[J]. 中国科技翻译, 2020, 33(4): 21-23+42.

[11] 李梅, 朱锡明. 英汉机译错误分类及数据统计分析[J]. 上海理工大学学报(社会科学版), 2013, 35(3): 201-207.

[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.