Modern Linguistics 现代语言学, 2025, 13(5), 604-609
Published Online May 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.135514

Hans汉斯

多模态机器翻译发展现状研究

沈思娴

上海海事大学外国语学院，上海

收稿日期：2025年2月17日；录用日期：2025年5月15日；发布日期：2025年5月29日

摘　要

随着科学技术的高速发展，多模态机器翻译(Multimodal Machine Translation, MMT)作为新兴的计算机辅助翻译技术，日益受到关注。本文聚焦于多模态机器翻译的发展现状，深入剖析其定义、技术路径、应用领域以及面临的挑战，并探讨其对翻译变革的影响。通过对相关文献的梳理和技术方法的分析，全面呈现多模态机器翻译的研究进展和实际应用场景，同时针对其面临的挑战提出相应的解决策略，旨在为该领域的进一步研究提供坚实的理论基础和有益的参考，推动多模态机器翻译技术的持续发展和在翻译实践中的更广泛应用。

关键词

多模态机器翻译，机器翻译，翻译技术

Research on Current Development Status of Multimodal Machine Translation

Sixian Shen

School of Foreign Languages, Shanghai Maritime University, Shanghai

Received: Feb. 17th, 2025; accepted: May 15th, 2025; published: May 29th, 2025

Abstract

The rapid advancement of science and technology has propelled Multimodal Machine Translation (MMT) into the spotlight as a cutting-edge computer-aided translation technology. This paper centers on the current development status of MMT, delving into its definition, technical pathways, application domains, challenges, and implications for translation transformation. By meticulously examining relevant literature and technical approaches, it offers a comprehensive overview of MMT's research progress and real-world applications. Moreover, it offers targeted solutions to the encountered challenges and contemplates the impact of MMT on the translation landscape, thereby laying

文章引用: 沈思娴. 多模态机器翻译发展现状研究[J]. 现代语言学, 2025, 13(5): 604-609.
DOI: 10.12677/ml.2025.135514

a robust theoretical groundwork for further exploration and propelling the ongoing evolution and practical deployment of MMT technology within the translation realm.

Keywords

Multimodal Machine Translation, Machine Translation, Translation Technology

1. 引言

在传统翻译理论中，翻译涉及将一种语言转录为另一种语言。许钧指出：“翻译是一种跨文化交际活动，其任务是通过符号转换再现意义”[1]。此外，语言学家卡特福德(Catford)认为，翻译是用另一种语言(目标语言，TL)替换一种语言(源语言，SL)中的等值文本材料[2]。根据这些对翻译的解释，可以初步推断，翻译涉及传递符号意义或文本语言转换的过程。

随后，随着语言处理技术的出现，翻译与人工智能相结合，机器翻译(Machine Translation, MT)成为新的发展方向。机器翻译指的是自动翻译或自然语言处理，包括利用计算机技术将一种自然语言翻译为另一种自然语言的过程，通常涉及自然语言句子和完整文本的翻译[3]。机器翻译在一定程度上能够辅助译者的日常工作或学习，提高翻译的效率。然而，随着人类社会与科技的进步，人际交流已不再仅限于传统的语言符号学模式。在现代社会，交流可以通过多种方式实现，这些方式由语言、图像、声音、色彩和动作等多种符号意象资源构成的复杂媒介所体现[4]，被认为是多模态媒介。因此，“多媒体性成为社会文化实践的基本模式，相应地，多模态性成为意义建构与互动的共同特征”[5]。我们可以将多模态性理解为通过一种以上的交流模式来传递信息，包括书面语言、口头语言、手势、声音、视觉图像等[6]。

事实上，许多需要翻译的文献或产品本身就是多模态的，由不同的相关模态构成。如今，为满足各个领域对多模态交互的多元化需求，多模态翻译已被广泛应用于多个场景，如网站、社交媒体和视听应用等。

2. 文献综述

多模态机器翻译(MMT)作为机器翻译领域的一个新兴方向，近年来受到了国内外研究者的广泛关注，相关研究已取得显著进展，但也存在一些有待进一步探讨的问题。

从研究方法来看，早期的研究多集中于将神经机器翻译与神经图像描述方法相结合，如 Desmond Elliot 等人的工作[7]，这种方法在一定程度上实现了多模态信息的融合，但融合方式较为简单，未能充分发挥多模态的优势。随后，研究者们开始探索更为复杂的多模态注意力机制，如 Stella Frank 等人提出的在神经机器翻译解码器中融入多种多模态注意力机制的模型[8]，这种方法在一定程度上提高了翻译的准确性和对多模态信息的利用效率，但模型的复杂度也相应增加，对计算资源的要求更高。近年来，研究更加注重对模型的精调和优化策略。例如，Ozan Caglayan 等人整合了两种方法来学习视觉支持的跨语言表示，并发现当对 MMT 进行精调时，各种模态能够实现最佳性能[9]。这表明，通过对模型进行适当的精调和优化，可以更好地发挥多模态的优势，提高翻译质量。

在应用场景方面，多模态机器翻译的应用范围不断扩大。早期的研究主要集中在图像与文本的结合，如多语言图像描述等任务。随着技术的发展，多模态机器翻译在旅游、外语学习、专业阅读、跨境电商

等多个领域展现出广泛的应用前景。例如，在跨境电商领域，多模态机器翻译能够实时翻译商品描述、用户评价、客服对话等，提升用户体验和平台运营效率。

然而，目前的研究仍存在一些不足之处。首先，多模态数据的获取和标注成本较高，构建高质量的多模态数据集是成功的关键，但目前公开可用的多模态数据集相对较少，这在一定程度上限制了多模态机器翻译技术的发展。其次，多模态融合策略仍有待进一步优化。虽然已经提出了多种融合方法，但在实际应用中，如何根据不同的任务和数据特点选择合适的融合策略，仍然是一个需要深入研究的问题。此外，多模态机器翻译的模型性能评估也是一个亟待解决的问题。目前，大多数研究主要采用传统的机器翻译评估指标，如 BLEU、TER 等，这些指标在评估多模态翻译结果时可能存在一定的局限性。因此，需要开发更加适合多模态翻译的评估指标和方法，以更全面、准确地衡量翻译质量。

综上所述，多模态机器翻译领域在过去 10 年内取得了较为活跃的研究成果，但仍有诸多问题需要进一步研究和解决。未来的研究应重点关注多模态数据的获取与标注、融合策略的优化以及模型性能评估等方面，以推动多模态机器翻译技术的进一步发展和应用。

3. 多模态机器翻译发展现状

3.1. 定义

在多模态机器翻译(MMT)过程中，“针对某一模态中的实体，其任务是在另一种模态中生成相同的实体”[10]。例如，输入一张图像后，MMT 可以生成描述该图像的句子；而输入文本描述，则可以提供相应的图像。目前，在涉及图像与文本模态的多模态机器翻译(MMT)中，支持这种双模态翻译活动的方式主要有两种。第一种方法旨在基于注意力机制连接文本与图像。在该方法中，双注意力解码器(dual-attentive decoder)能够自然地整合通过预训练卷积神经网络(CNN)提取的空间视觉特征，从而缩小图像描述与翻译之间的差距。在翻译生成过程中，双注意力解码器能够分别处理图像块和源语言文本。第二种方法主要侧重于将图像作为 MMT 模型的输出部分之一。换言之，预训练卷积神经网络提取的视觉特征被融入基于注意力机制的神经机器翻译模型，使其成为翻译过程中的辅助信息。此外，为了进一步提升图像-文本交互模态翻译的效果，研究者尝试将多模态翻译分解为两个子任务——“学习翻译”(learning to translate)与“学习视觉支持表示”(learning visually grounded representations)。在他们构想的模型中，通过训练，翻译机器可以预测与源语言相关的视觉特征，从而增强翻译的语义关联性和准确性。

3.2. 多模态机器翻译应用

借助多模态翻译或多模态性，人们探索了将多模态翻译与其他传统中英平行语料库相结合的不同方式。在一定程度上，这类应用能够促进相关翻译体验或学习的提升。

3.2.1. 多模态旅游翻译语料库

由于多模态语料库在不同领域的研究价值，近年来受到了众多学者的关注。同时，国内研究者也致力于构建中英平行翻译语料库。例如，胡开宝教授带领团队成功建立了智能化多语种教学与研究平台，该平台涵盖了政府工作报告中英平行语料库、莎士比亚戏剧中英平行语料库以及新闻发布会口译中英语料库。然而，“这些语料库大多局限于文本层面，缺乏多模态语料库及数字技术的支持”[11]。

为了促进旅游翻译的发展以及旅游文化的海外传播，胡富茂等人[11]通过收集文本、音频、视频和图像语料并进行处理，构建了多模态旅游翻译语料库。该语料库由四种模态(文本、图像、音频和视频)组成，主要包括洛阳旅游文化中英平行语料库和八大古都中英平行语料库。

由于多模态旅游翻译语料库包含了文本、图像、音频和视频等标注材料，并可通过关键词检索，当

用户输入“中国的首都”时，语料库能够提供匹配的原始中文文本及其英文翻译“Beijing is the capital of China”。此外，系统还会同步显示与“北京”“中国”“首都”相关的音频、视频及图像，以补充文本检索结果。因此，凭借多模态特性，该语料库使游客能够从多维度全面获取和感知旅游景点的相关信息，从而提升其旅游体验。

3.2.2. 多模态口译语料库

口译语料库是口译研究的重要工具，相关研究已取得丰富成果。然而，目前的口译语料库主要是由录音转写而成的单模态文本语料库[12]。尽管这类语料库在一定程度上可用于口译研究，但由于在转写过程中丢失了语音、口译场景及手势、表情等视觉信息，它无法全面反映口译的真实情境与过程。因此，现有的单模态文本口译语料库存在一定局限性。

针对这一问题，刘剑与胡开宝提出构建多模态口译语料库。在该语料库中，用户在点击任何检索结果后，相应的音频或视频(包括讲话语音、表情和手势)可同步呈现，使口译学习者能够直观观察口译的真实过程，从而获得更优的学习体验。

4. 当前 MMT 所面临的挑战与解决

4.1. 面临的挑战

(1) 多模态信息融合难题

目前，多模态机器翻译面临的一个关键问题是缺乏通用的多模态信息融合框架[13]。在图像–文本机器翻译和视频–文本机器翻译这两个主要子任务中，尚未形成一个统一且有效的多模态融合框架，这使得不同模态之间的信息整合存在困难，不利于翻译质量的提升。例如，在处理复杂的多模态输入时，模型可能难以准确判断哪些视觉信息与文本信息相关，以及如何将它们有机结合起来进行翻译，从而导致翻译结果不准确或不完整。

(2) 无关信息的干扰

部分图像中可能包含与文本无关的信息，这些无关的图像信息会干扰机器翻译模型，从而影响翻译结果，甚至导致模型生成错误的翻译内容[14]。例如，在翻译一篇包含人物照片和相关文字介绍的文章时，照片中人物的背景装饰等与文章主题无关的视觉信息可能会误导模型，使其在翻译时过度关注这些无关细节，而忽略了文本的核心内容，进而产生不准确的翻译。

(3) 视频时间特征的处理不足

与图像不同，视频与文本具有时间顺序特性，视频中每一帧的特征均具有序列性。然而，当前的多模态机器翻译模型在处理视频信息时，往往未充分考虑视频中不同时间特征对文本的影响。对于某个特定的词语，视频在不同时刻的语义关联程度存在显著差异，这种时间维度上的语义变化难以被准确捕捉和利用，从而限制了模型在视频–文本翻译任务中的表现。

4.2. 解决策略

(1) 构建统一的多模态融合框架

为了克服多模态信息融合的难题，研究者需要致力于构建一个通用且高效的多模态融合框架[15]。该框架应能够适应不同类型的多模态输入，并有效地整合文本、图像、视频等多种模态的信息。例如，可以探索基于深度学习的融合方法，通过设计特定的神经网络架构，使模型能够自动学习不同模态之间的关联和映射关系，实现信息的无缝融合。同时，结合注意力机制等技术，让模型在翻译过程中能够更加关注与文本语义紧密相关的视觉或听觉信息，提高翻译的准确性和连贯性。

(2) 信息筛选与相关性分析

针对无关信息的干扰问题，可以在多模态机器翻译模型中引入信息筛选和相关性分析模块[13]。在翻译前，对输入的多模态信息进行预处理，通过分析文本和图像、视频等之间的语义相关性，筛选出与文本主题高度相关的视觉或听觉信息，同时剔除无关或干扰信息。例如，利用图像识别技术和文本语义理解技术，判断图像中的哪些元素与文本内容直接相关，只保留这些相关信息用于翻译，从而减少无关信息对模型的干扰，提升翻译质量。

(3) 强化视频时间特征的处理

对于视频时间特征处理不足的问题，需要进一步改进多模态机器翻译模型对视频信息的处理机制[14]。可以采用时序编码器等技术，对视频的每一帧进行时间序列上的编码，捕捉视频在不同时间点的语义变化和特征演化。同时，将这些时间特征与文本语义进行关联分析，使模型能够根据视频的时间特性更准确地理解文本含义，并生成符合语境的翻译结果。例如，在翻译电影字幕时，模型可以结合视频画面的时间流动和情节发展，生成更贴合实际场景的字幕翻译，增强观众的理解和观看体验。

5. 总结

随着科学技术的飞速发展，多模态机器翻译(MMT)作为机器翻译领域的重要发展方向，已取得了显著的研究进展并在多个应用领域展现出巨大潜力[16]。本文系统地探讨了多模态机器翻译的定义、技术路径、应用领域、面临的挑战以及对翻译变革的影响，全面呈现了其发展现状。在技术路径方面，基于注意力机制的图文合成方法、图像作为模型输出的翻译方法以及多模态翻译的子任务分解方法等，为 MMT 的实现提供了有效的技术支持，使其能够更好地整合多模态信息，提升翻译质量。在应用领域，多模态旅游翻译语料库和多模态口译语料库等实际应用案例，充分展示了 MMT 在促进旅游文化传播和口译学习等方面的独特优势，为相关行业的创新发展提供了有力支持。然而，MMT 在发展过程中也面临着多模态信息融合难题、无关信息干扰以及视频时间特征处理不足等挑战，针对这些问题，本文提出了构建统一的多模态融合框架、信息筛选与相关性分析以及强化视频时间特征处理等解决策略，以期为未来的研究和实践提供有益的参考。总之，多模态机器翻译的不断进步将为翻译行业带来更多的机遇和变革，有望在更广泛的领域实现更高效、更准确的跨语言交流，推动全球信息共享和文化交流的深入发展。

参考文献

[1] 许钧. 翻译概论(修订版) [M]. 北京: 外语教学与研究出版社, 2020.

[2] Catford, J. (1965) A Linguistics Theory of Translation. Oxford University Press.

[3] Li, L. (2014) Evolving Academic Libraries in the Future. In: Li, L., Ed., *Scholarly Information Discovery in the Networked Academic Learning Environment*, Elsevier, 279-309. https://doi.org/10.1533/9781780634449.4.279

[4] 李小华, 唐青叶. 国内多模态翻译研究的可视化分析: 现状、问题及建议[J]. 北京科技大学学报(社会科学版), 2021, 37(5): 534-542.

[5] 吴赟. 媒介转向下的多模态翻译研究[J]. 外国语(上海外国语大学学报), 2021, 44(1): 115-123.

[6] Díaz-Cintas, J. and Muñoz-Sánchez, P. (2006) Fansubs: Audiovisual Translation in an Amateur Environment. *Journal of Specialised Translation*, No. 6, 37-52.

[7] Elliott, D., Frank, S. and Hasler, E. (2016) Multilingual Image Description with Neural Sequence Models. *Proceedings of the International Conference on Learning Representations*, 1, 1-16.

[8] Calixto, I., Elliott, D. and Frank, S. (2016) DCU-UvA Multimodal MT System Report. *Proceedings of the First Conference on Machine Translation*, **2**, 634-638. https://doi.org/10.18653/v1/w16-2359

[9] Caglayan, O., Kuyu, M., Amac, M.S., Madhyastha, P., Erdem, E., Erdem, A., *et al*. (2021) Cross-Lingual Visual Pre-Training for Multimodal Machine Translation. *Proceedings of the* 16*th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume*, 19-23 April 2021, 1317-1324. https://doi.org/10.18653/v1/2021.eacl-main.112

[10] Baltrusaitis, T., Ahuja, C. and Morency, L. (2019) Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 423-443. https://doi.org/10.1109/tpami.2018.2798607

[11] 胡富茂, 宋江文, 王文静. 多模态旅游翻译语料库建设与应用研究[J]. 上海翻译, 2022(5): 26-31.

[12] Calixto, I., Liu, Q. and Campbell, N. (2017) Doubly-Attentive Decoder for Multi-Modal Neural Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vancouver, 30 July-4 August 2017, 1913-1924. https://doi.org/10.18653/v1/p17-1175

[13] Caglayan, O., *et al*. (2020) What Do You Mean? Visually-Grounded Meaning Representations for Multimodal Machine Translation. arXiv:2001.08999.

[14] Specia, L., *et al*. (2020) A Survey of Multimodal Translation. arXiv:2005.12139.

[15] Specia, L., *et al*. (2020) Findings of the WMT 2020 Shared Task on Multimodal Machine Translation. arXiv:2010.07652.

[16] 黄鑫, 张家俊, 宗成庆. 基于跨模态实体信息融合的神经机器翻译方法[J]. 自动化学报, 2023, 49(1): 1-11.