

文章编号: 1003-0077(2025)09-0001-22

实时翻译研究综述

张绍磊^{1,3}, 冯 洋^{1,2,3}

(1.中国科学院 计算技术研究所 智能信息处理重点实验室, 北京 100190;

2.中国科学院 智能算法安全重点实验室, 北京 100190;

3.中国科学院大学, 北京 100049)

摘 要: 标准的机器翻译通常需要等待全部源语言输入后再开始翻译, 这种离线交互模式难以满足同声传译、国际会议、跨国旅游等实时场景中的翻译需求。与之相比, 实时翻译要求模型在接收源语言输入的同时生成目标语言输出, 从而在低延时下完成翻译, 因此逐渐受到研究人员的广泛关注。然而, 由于需要决策何时开始翻译并基于不完整的输入进行翻译, 实时翻译在性能上面临诸多挑战。近年来, 许多研究致力于构建低延时高质量的实时翻译模型, 但仍然缺少对于现有方法和研究趋势的系统梳理。该文不仅从不同应用场景出发, 对现有实时翻译方法进行了详细分类和总结分析, 还对实时翻译研究的发展趋势进行了展望, 并系统性地整理了相关文献, 为研究人员提供了快速了解该领域最新动态和未来发展方向的参考。

关键词: 实时翻译; 机器翻译; 语音翻译

中图分类号: TP391

文献标识码: A

A Survey of Simultaneous Translation

ZHANG Shaolei^{1,3}, FENG Yang^{1,2,3}

(1.Key Laboratory of Intelligent Information Processing,

Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS), Beijing 100190, China;

2.Key Laboratory of AI Safety, Chinese Academy of Sciences, Beijing 100190;

3.University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Standard machine translation typically requires waiting for the complete source language input before beginning the translation. This offline interaction model falls short in meeting the demands of real-time scenarios, such as simultaneous interpretation, international conferences, and national travel. In contrast, simultaneous translation necessitates that the model generates target outputs while receiving the source inputs, thereby achieving translation under low latency. As a result, it has garnered increasing attention from researchers. However, simultaneous translation faces several challenges, primarily due to the need to determine the appropriate moment to begin translating and to generate translations based on incomplete inputs. Recently, much research has focused on developing low-latency, high-quality simultaneous translation models, yet there remains a lack of a comprehensive review of existing methods and research trends. This paper not only provides a detailed classification and analysis of current simultaneous translation methods from the perspective of application scenarios but also forecasts future research directions. Furthermore, this paper systematically organizes relevant literature, offering researchers a valuable resource to quickly grasp the latest advancements and emerging trends in the field of simultaneous translation.

Keywords: simultaneous translation; machine translation; speech translation

收稿日期: 2025-02-11 定稿日期: 2025-04-02

基金项目: 国家自然科学基金(62376260)

0 引言

翻译作为自然语言处理领域的重要分支,旨在将一种语言(源语言)自动转化为另一种语言(目标语言),在跨语言交流中扮演着至关重要的角色。随着机器翻译技术的飞速发展,传统的离线翻译系统已能够提供高质量的译文。然而,传统离线翻译系统需要等待源语言全部输入完成后才能开始翻译,这使其难以应对如国际会议、跨国旅游和在线直播等实时场景中的翻译需求。随着实时场景下翻译需求的不断增加,研究者们逐渐将目光转向实时翻译技术,致力于在低延时下提供高质量的翻译结果。2014 年,Grissom II 等人^[1]首次提出了实时翻译(Simultaneous Translation)这一概念,旨在通过接收源语言(语音或文本)输入的同时,实时地将其翻译成目标语言(语音或文本)并呈现给用户。此后,实时翻译的研究逐年增加。如图 1 所示,过去十年中每年公开的实时翻译相关研究数量逐年增长,实时翻译逐步成为机器翻译领域的研究热点。

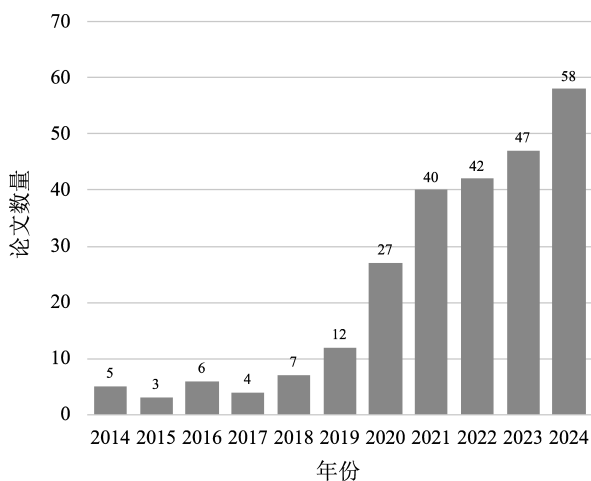


图 1 近十年实时翻译研究的公开论文数量

在过去几年中,实时翻译的研究经历了四个发展阶段:文本到文本实时翻译^[2]、语音到文本实时翻译^[3]、语音到语音实时翻译^[4],以及基于大模型的实时翻译^[5],整体呈现出模型能力不断增强、架构逐步端到端化的趋势。在初期阶段,许多研究集中在文本到文本实时翻译,该类研究通常需要将语音识别模块与实时翻译模块级联,以实现跨语言沟通。具体而言,语音识别模块首先将说话者的语音转录为文本,随后文本到文本实时翻译模块将该文本实时翻译为目标语言文本。随着研究的深入,学者们逐渐转向端到端的语音到文本实时翻译,旨在直接

将源语言语音输入转化为目标语言文本输出,从而有效减少级联过程中的误差和延迟。更进一步,语音到语音实时翻译被提出,其目标是在接收源语言语音输入的同时,实时生成目标语言语音。这一端到端生成语音的特点,使得在实时翻译中保留音色和韵律成为可能。近年来,随着大模型技术的兴起,基于大模型的实时翻译研究开始探索如何利用大模型的生成能力,进一步提升实时翻译的质量。

目前,实时翻译领域已有大量研究成果,主要集中在构建低延迟、高质量的实时翻译系统。然而,现有文献中尚缺乏对实时翻译方法进行全面梳理与总结的综述性文章。部分关于神经机器翻译的综述文章^[6-7]中,仅以简短的一节介绍实时翻译的前沿趋势;而另一部分聚焦于语音翻译的文章^[8]提及了实时翻译任务。现有研究缺乏对实时翻译的系统性综述,未对实时翻译任务在文本到文本、语音到文本、语音到语音以及基于大模型方法的发展进行全面总结。这四个阶段在发展上内在联系紧密,逐步递进,对于研究人员深入理解实时翻译技术及未来研究方向具有重要价值。本文系统性地综述实时翻译的任务定义、主要方法以及未来发展趋势,重点介绍现有的实时翻译方法,并总结不同方法之间的优缺点与内在联系。本文涵盖了最新的实时翻译研究成果,能够帮助研究人员了解该领域的前沿进展。为了帮助研究人员快速掌握实时翻译领域的最新进展,并为未来的研究提供参考和方向性指导,本文对实时翻译的相关研究文献进行了系统性整理,并按照时间、方法及会议进行了详细分类^①。

本文组织结构如下:第 1 节简要介绍实时翻译的背景;第 2 节总结分析实时翻译的研究进展并进行归类梳理;第 3 节展望实时翻译未来的研究方向;第 4 节总结全文。

1 背景

本节主要介绍机器翻译和实时翻译的基本概念,分析当前机器翻译与实时翻译的特点及其面临的问题,最后阐述实时翻译所面临的关键挑战。

1.1 机器翻译的概念

机器翻译(Machine Translation)指的是利用计算机技术自动将一种自然语言转换为另一种自然语言的过程。近年来,深度学习技术的发展大幅提高

^① 实时翻译的文献整理: <https://github.com/ictnlp/Simultaneous-Translation-Survey>

了机器翻译的性能,基于编码器-解码器(Encoder-Decoder)架构^[9]的神经机器翻译已成为主流的翻译框架。机器翻译的训练数据通常由平行语料组成,表示为 $D = \{(X, Y)\}$, 其中, $X = (x_1, \dots, x_m)$ 为源语言输入序列, $Y = (y_1, \dots, y_n)$ 为目标语言输出序列。通过这些平行数据,模型能够建模从源语言序列到目标语言序列之间的映射关系 $P_{mt}(Y|X)$, 其概率表示如式(1)所示。

$$P_{mt}(Y|X) = \prod_{i=1}^n P(y_i | y_{<i}, X) \quad (1)$$

其中, y_i 表示当前输出的目标词, $y_{<i}$ 表示之前输出的所有目标标记。

在机器翻译中,每个目标词 y_i 的生成都依赖于完整的输入序列 X , 如图 2(a) 所示。因此,标准的机器翻译模型通常采用离线方式处理输入,即在等待源语言输入完全结束后,才开始翻译过程。这种处理方式导致用户必须等到源语言输入完全结束,才能获得翻译结果,从而产生较高的延时。因此,离线机器翻译在同声传译、实时字幕等对实时性要求较高的应用场景中难以满足需求。

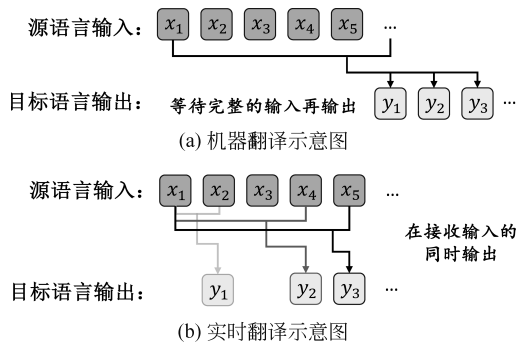


图 2 实时翻译和机器翻译的对比示意图

1.2 实时翻译的概念

实时翻译(Simultaneous Translation)是指在接收源语言输入的同时,实时将其转换为目标语言输出的技术^[1]。与离线机器翻译不同,实时翻译无须等待完整的源语言输入结束才能开始翻译,而是在接收源语言的过程中便能同步生成目标语言输出。这使得实时翻译能够在低延迟的条件下向用户呈现翻译结果,特别是在同声传译、实时字幕等对实时性要求极高的应用场景中具有重要价值。根据输入模态和输出模态(文本或语音)的不同,实时翻译模型可以进一步细分为文本到文本、语音到文本、语音到语音等具体任务。

在实时翻译过程中,模型不再依赖完整的源语言序列来生成每个目标标记,而是仅基于当前接收到的部分源语言输入生成对应的目标标记。因此,

实时翻译需要额外设计实时策略,以决定模型是继续等待更多的源语言输入还是开始生成翻译。具体而言,图 2(b)展示了实时翻译的过程示例:模型在接收到 x_1 和 x_2 之后立即生成 y_1 ,在继续接收到 x_3 和 x_4 之后立即生成 y_2 ,然后在接收到 x_5 之后立即生成 y_3 ,以此类推。决策何时生成每个目标标记即为实时策略,是实时翻译的关键挑战之一。形式化地,用 $g(i)$ 表示生成 y_i 的时机,即在接收到前 $g(i)$ 个输入时生成 y_i 。从源语言输入序列翻译为目标语言输出序列的概率表示如式(2)所示。

$$P_{st}(Y|X) = \prod_{i=1}^n P(y_i | y_{<i}, x_{<g(i)}) \quad (2)$$

其中, $x_{<g(i)}$ 表示前 $g(i)$ 个源语言输入标记。

实时翻译通过基于部分输入进行翻译,能够在低延迟下实时呈现翻译结果。然而,由于模型仅能依赖当前接收到的输入进行翻译,其可用的输入信息是有限的,这将直接影响翻译质量。因此,如何合理地确定 $g(i)$,即在何时翻译每个目标标记 y_i ,成为实时翻译中的关键问题。

1.3 实时翻译的评价指标

与离线机器翻译不同,实时翻译要求模型在低延迟下生成高质量的翻译。因此,实时翻译通常采用翻译质量和延时两个评价指标进行双维度评估,以衡量在一定延时下的翻译效果。

翻译质量 实时翻译沿用离线机器翻译的指标来评估翻译质量,包括基于统计的 BLEU 值和基于语义相似度的 COMET 值等。翻译质量越高,意味着实时翻译模型的性能越好。

延时 平均滞后^[2] (Average Lagging, AL) 是实时翻译中最常用的延时评价指标,其评估实时翻译模型的输出相对于输入的平均滞后量。对于实时翻译过程,根据翻译 y_i 的时机 $g(i)$, AL 的具体计算为如式(3)、式(4)所示。

$$AL = \frac{1}{\tau} \sum_{i=1}^{\tau} g(i) - \frac{i-1}{|Y| / |X|} \quad (3)$$

$$\tau = \min\{i \mid g(i) = |X|\} \quad (4)$$

其中, $|X|$ 和 $|Y|$ 分别表示输入序列和输出序列的长度, τ 为源语言序列结束输入的时机。延时越低,意味着实时翻译模型的性能越好。

近来,长度自适应平均滞后 (Length Adaptive Average Lagging, LAAL)^[10] 被提出,其是平均滞后 (AL) 的一种改进版本,同时考虑了过度生成 (Over-generation) 现象。LAAL 通过将话语长度除以参考文本长度和模型预测长度两者之间的最大值来定义理想延迟。当预测较短 (欠生成, Under-

generation)时,延时计算使用参考文本长度;当预测较长(过度生成)时,延时计算使用预测文本长度。具体而言,在 LAAL 的具体计算中将式(3)改为式(5)。

$$AL = \frac{1}{\tau} \sum_{i=1}^{\tau} g(i) - \frac{i-1}{\max(|Y|, |Y^*|)/|X|} \quad (5)$$

其中, $|Y^*|$ 为生成译文 Y^* 的长度。

1.4 实时翻译面临的挑战

根据上述概念,实时翻译模型需要在接收到源语言输入的同时确定何时开始翻译,并且只能依赖当前已接收到的源语言信息生成目标语言输出。这一实时性限制显著增加了建模与模型训练的难度。因此,与传统的机器翻译相比,实时翻译面临以下主要挑战。

(1) **实时策略: 如何确定何时开始翻译** 实时翻译的核心挑战之一在于如何在接收到部分源语言信息后,准确选择合适的时机开始翻译。如果翻译的启动时机不准确,可能导致翻译质量的下降或延时的增加。为了解决这一问题,实时翻译模型需要制定精确的策略,以控制翻译的开始时机,确保在低延时条件下生成高质量的翻译。

(2) **翻译能力: 如何基于不完整的源序列进行翻译** 传统机器翻译依赖完整的源语言序列进行翻译,而实时翻译则需要在源语言输入不完整的情况下完成翻译任务。这一特性可能导致翻译质量下降,尤其是在需要准确理解上下文的场景中。因此,实时翻译模型需要在训练过程中加强基于不完整输入进行翻译的能力,确保即使部分输入缺失,翻译结果依然具备连贯性和准确性。

(3) **端到端架构: 如何端到端地完成实时翻译** 早期的实时翻译模型通常采用级联式架构,聚焦于文本到文本的实时翻译,并结合外部模块如语音识别等。随着端到端模型的兴起,越来越多的研究探索如何通过端到端的方式实现语音到文本的实时翻译。与文本输入不同,语音信号通常是连续的且具有较大多样性,这使得在接收语音输入的同时,准确找到翻译时机并生成高质量翻译结果变得更加具有挑战性。最近的研究开始关注语音到语音的实时翻译,其端到端建模面临的挑战进一步加剧。

2 实时翻译方法分类

自实时翻译概念提出以来,近年来各种实时翻译模型不断涌现,旨在在低延时下实现高质量的翻译。根据具体应用场景和核心挑战的不同,现有的

实时翻译方法可分为以下四类: ① 文本到文本实时翻译方法; ② 语音到文本实时翻译方法; ③ 语音到语音实时翻译方法; ④ 基于大模型的实时翻译方法,如图 3 所示。这 4 类方法呈递进关系,并且具有紧密的内在联系。接下来,本节将详细介绍上述方法,并进一步分析和探讨当前的研究及技术进展。

2.1 文本到文本实时翻译方法

在实时翻译概念提出之初,大量研究集中于文本到文本实时翻译方法,该方法将源语言文本实时地转化为目标语言文本。文本到文本实时翻译的研究主要从 4 个主要方向展开: 固定策略、自适应策略、训练方法和数据增强,旨在在低延时下实现高质量的翻译效果。

固定策略和自适应策略的研究侧重于设计精确的实时翻译策略,以确定模型在适当时机开始翻译,从而确保在低延时的情况下维持高质量的翻译。训练方法的研究则聚焦于通过优化训练过程,提升模型在源语言信息不完整情况下的翻译能力,进而应对实时翻译中可能遇到的源语言序列不完整的挑战。数据增强方面的研究则通过对训练数据进行扩展或改写,提高数据质量,使其更加符合实时翻译的需求与偏好。本节将详细介绍并探讨上述 4 个研究方向。

2.1.1 固定策略

固定策略是早期出现的实时翻译方法,其通过预先制定的规则控制模型执行实时翻译。Cho 等人^[11]提出了 Wait-If-Worse 和 Wait-If-Diff 两种策略,基于启发式原则通过生成目标词的概率来控制模型是等待输入还是开始翻译。如果下一个目标词的概率降低或目标词发生变化,则模型选择等待新的输入;反之,则开始翻译。由于这两种策略涉及推理过程,因此仅能在推理阶段应用于离线机器翻译模型。Dalvi 等人^[12]提出了 STATIC-RW 策略,要求实时翻译模型交替翻译 RW 个目标语言标记,并等待 RW 个源语言标记,直到翻译完成。在 STATIC-RW 的基础上,Ma 等人^[2]借鉴了人类同声传译员“边听边翻译”的方式,提出 Wait- k 策略。该策略要求模型首先等待 k 个源语言标记输入,然后开始翻译,每接收到一个源语言标记便生成一个目标语言标记,保持输入与输出之间始终存在 k 个标记的差距。 k 为设定的超参数,通过调整 k 值可以控制模型在不同延时下进行翻译,从而平衡实时性与翻译质量。形式化地,Wait- k 策略中的翻译时机 $g(i)$ 可表示为 $g(i) = i + k - 1$ 。由于 Wait- k 策略具有简单性和可控性,所以它迅速成了应用最广

泛的实时翻译方法之一,并为实时翻译模型研究提供了重要的基础。

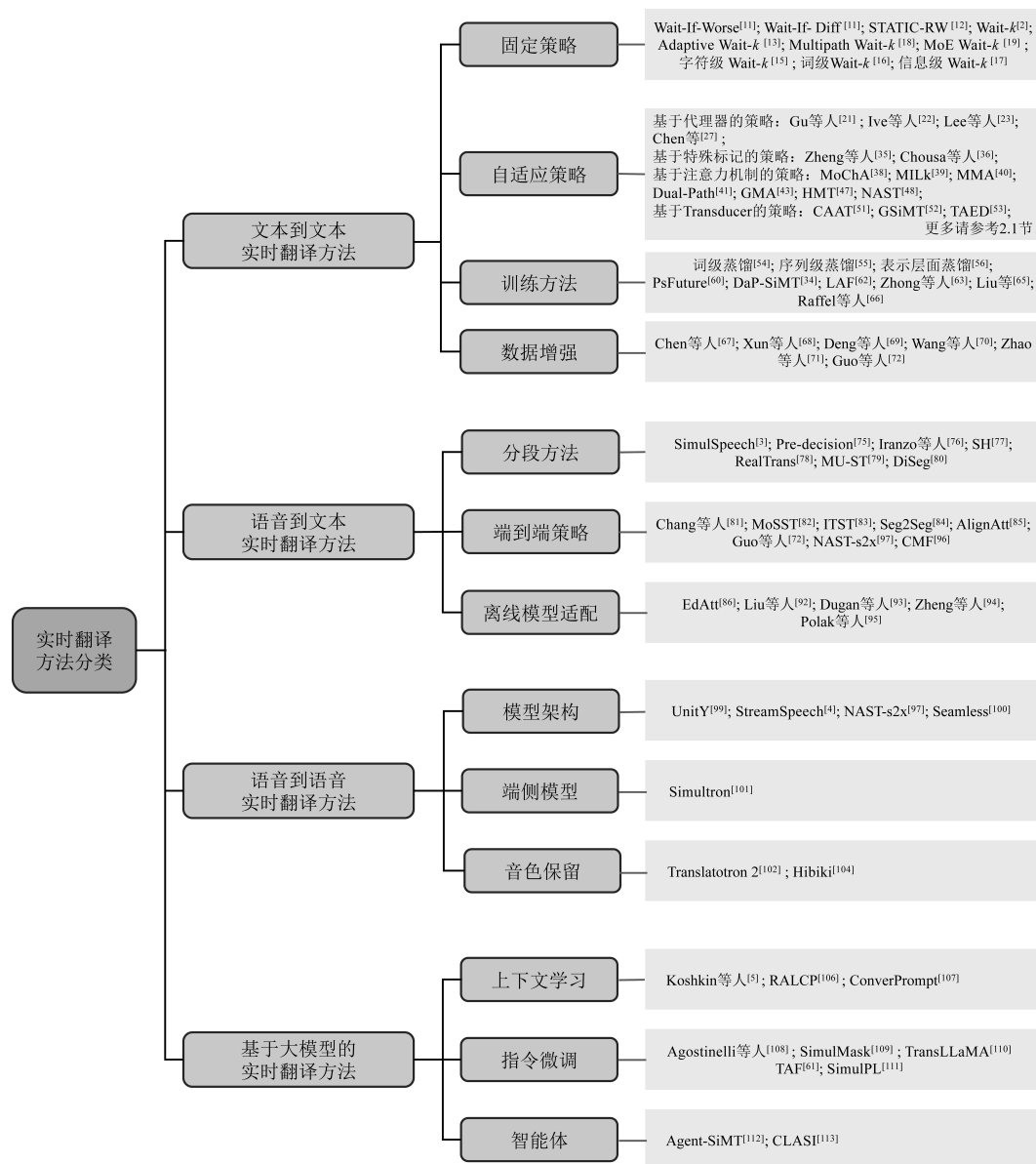


图3 实时翻译方法分类

尽管 Wait- k 策略已能够提供较为理想的实时翻译性能,但仍存在两方面的局限性。首先,固定的 Wait- k 策略忽视了上下文信息,模型可能在源语言信息尚不完整时被强制提前生成目标语言标记,进而影响翻译质量。其次,Wait- k 策略通过调整等待时间 k 来适应不同延时需求,但该策略需要为每种延时训练并维护多个 Wait- k 模型(如 Wait-1、Wait-3、Wait-5 等多个模型),从而增加了训练与部署的开销。为突破这两方面的局限性,后续研究提出了相关改进方法。

实时策略改进 为了解决 Wait- k 策略未能充

分利用上下文信息的问题,Zheng 等人^[13]提出了 Adaptive Wait- k 策略,通过集成多个 Wait- k 模型克服了单一 Wait- k 策略的限制。具体而言,该方法维持多个不同的 Wait- k 模型,根据生成目标语言标记的概率来动态选择合适的 Wait- k 模型。当目标标记生成的概率低于设定的最低阈值时,表明当前源语言输入不足以产生高质量的翻译,Adaptive Wait- k 会继续等待输入,并切换至 k 值更大的 Wait- k 模型;反之,当目标标记的生成概率超过设定的最高阈值时,说明源语言信息已足够,此时模型会切换至 k 值较小的 Wait- k 模型继续翻译。通过

这种方式, Adaptive Wait- k 能够根据上下文信息动态调整翻译策略, 从而提升实时翻译的性能。Guo 等人^[14]在 Wait- k 策略的基础上引入了后评估模块, 评估当前 Wait- k 策略产生的翻译质量, 并根据评估结果调整 Wait- k 策略的滞后程度(即 k 值)。这种方法不仅能动态适应翻译需求, 还能有效优化翻译质量。此外, 针对不同语言的特点, 一些研究提出了不同粒度的 Wait- k 策略(如字符级^[15]、词级^[16]、信息级^[17]), 以便分别处理中文、日文、德文等特定语言, 取得了不错的性能提升。

通用性提升 为了解决 Wait- k 策略需训练并维护多个模型的问题, Elbayad 等人^[18]提出了 Multipath Wait- k 策略。该方法通过在训练过程中随机采样不同的滞后标记数 k , 使得模型能够在多个延时下进行实时翻译。该方法仅需一个模型便可适应多种延时需求, 从而有效减少了模型数量, 降低了训练与部署的成本。进一步地, Zhang 等人^[19]提出了 MoE Wait- k 策略, 通过混合专家(Mixture-of-Experts)构建支持多个延时的通用实时翻译模型。MoE Wait- k 包含多个专家模块, 每个专家负责特定延时下的实时翻译, 并协同决定最终的翻译结果。MoE Wait- k 仅需一个模型, 即在所有延时下获得

当前固定策略中的最佳表现。Multipath Wait- k 与 MoE Wait- k 策略不仅通过减少所需模型数量降低了训练与部署的开销, 还通过在训练过程中引入不同延时的固定策略相互促进, 进而提升不同延时下的翻译性能。

2.1.2 自适应策略

自适应策略根据当前接收到的输入动态决定是生成目标标记还是等待更多输入信息, 从而有效提高实时翻译性能。根据其建模方式, 自适应策略可以分为 4 类: 基于代理器的自适应策略、基于特殊标记的自适应策略、基于注意力机制的自适应策略以及基于 Transducer 的自适应策略。

基于代理器的自适应策略 早期的自适应策略通过在翻译模型中引入额外的代理器, 使其能够根据当前上下文状态动态输出两种操作: 读操作(READ)和写操作(WRITE), 如图 4(a)所示。其中, 读操作控制模型等待源语言标记的输入, 而写操作则控制模型生成目标语言标记。这类方法将实时翻译策略的学习转化为对读/写操作的预测, 相关研究通过强化学习、监督学习以及模仿学习等方式训练代理器, 使其能够在不同上下文条件下动态调整模型的读/写操作。

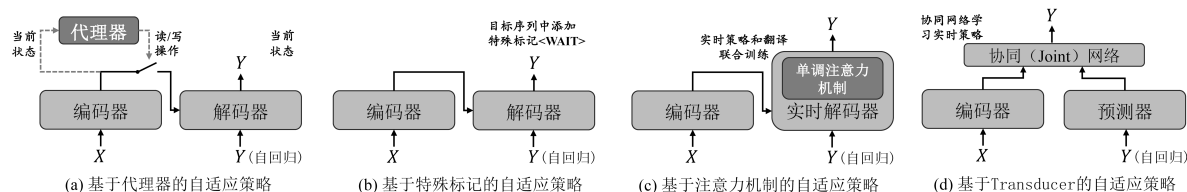


图 4 三类自适应策略的示意图

强化学习最早被用来训练代理器^[1,20-21]。在每个时刻, 代理器根据当前上下文生成读/写操作, 翻译模型进行翻译。翻译模型和代理器的训练分别进行, 其中翻译模型通过标准的机器翻译任务进行训练, 而控制操作的代理器则通过强化学习进行训练, 其中奖励函数综合考虑了翻译质量和延迟两个因素。通过强化学习训练得到的代理器能够有效控制翻译模型的读/写操作, 成为早期主流的实时翻译方法之一。在此基础上, Ive 等人^[22]和 Lee 等人^[23]等进一步将基于强化学习的实时翻译策略扩展至引入更多上下文信息、视觉输入和注意力机制, 以提升代理器在预测读/写操作时的准确性。

随后, 研究者们趋向于采用更加简化的监督学

习框架, 通过直接使用读/写操作的 ground-truth 序列来对代理器进行监督训练。然而, 现有数据通常缺乏精确的 ground-truth 读/写操作, 这一问题促使研究者们集中精力构建更加准确的读/写操作以供训练使用。为解决这一难题, Zheng 等人^[24]提出了一个假设: 若模型预测目标词的准确率足够高, 则说明源语言信息已经充分, 模型可以执行写操作; 反之, 则应执行读操作。基于这一假设, 大量研究通过预先训练的离线机器翻译模型, 根据目标词生成的概率来构造 ground-truth 读/写操作, 并利用这些标注训练代理器^[24-26]。此外, Chen 等人^[27-28]根据模型生成停止标记<eos>的概率来构建 ground-truth 读/写操作, 并通过此序列训练代理器。另一种构建

ground-truth 读/写操作的方法是利用标准译文,许多研究通过将源语言序列逐步输入训练好的机器翻译模型,并检测生成的译文与标准译文的一致性来构建用于训练代理器的 ground-truth 读/写操作^[29-34]。如果生成的内容和标准译文一致,则表明源语言信息完整,模型可以开始翻译;若不一致,则需继续等待。

尽管基于代理器的自适应策略在实时翻译中性能较好,但代理器和翻译模型通常是独立训练的,这导致实时翻译策略的学习与翻译模型的学习相互割裂,从而难以实现最佳的实时翻译策略。

基于特殊标记的自适应策略 有些研究通过引入特殊标记将预测读/写操作建模为生成任务,如图 4(b)所示。具体而言,Zheng 等人^[35]和 Chousa 等人^[36]通过向目标词表中加入〈WAIT〉令牌来建模读操作。当模型生成〈WAIT〉时,模型等待新的输入;其余情况下模型翻译目标标记。生成〈WAIT〉令牌的过程通常采用模仿学习^[35](Imitation Learning)或连接时序分类^[36](Connectionist Temporal Classification, CTC)方法进行训练。然而,特殊标记的引入往往会增加训练复杂度,降低实时翻译模型的训练效率。

基于注意力机制的自适应策略 为实现翻译模型与实时翻译策略的联合训练,许多研究开始探索如何在翻译模型的编码器-解码器架构中有效融合实时翻译策略。作为控制目标序列与源序列之间关注关系的关键技术,注意力机制逐渐成为这一目标的首选方法。如图 4(c)所示,基于注意力机制的自适应策略通常通过设计新的交叉注意力机制来实现翻译与实时翻译策略的联合学习。

在实时翻译任务中,目标序列对源序列的关注通常表现为单调性。具体来说,目标序列中的每个目标标记仅能关注当前接收到的源标记,因此,目标序列中后续标记所能关注的源标记数量是递增的。这一特性促使 Raffel 等人^[37]首次提出了单调注意力(Monotonic Attention)的概念,并为后续的实时翻译研究奠定了基础。在生成目标标记 y_i 时,单调注意力按照从左到右的顺序处理输入序列 X 。对于每个源标记 x_j ,模型计算一个伯努利选择概率 $z_{i,j}$,若 $z_{i,j} > 0.5$,则模型在接收到 x_j 后生成 y_i ;否则,模型继续读取下一个源标记 x_{j+1} ,并计算 $z_{i,j+1}$ 。单调注意力通过期望方式将伯努利选择变

量融入交叉注意力机制的计算中,从而使得开始翻译或等待的决策(即 $z_{i,j}$)能与翻译模型联合训练。该方法为后续实时翻译研究提供了理论支持。在此基础上,Chiu 等人^[38]提出了 MoChA(Monotonic Chunkwise Attention),该方法在单调注意力的基础上增强了局部窗口内的注意力。Arivazhagan 等人^[39]则提出了 MILk(Monotonic Infinite Lookback Attention),在 LSTM 翻译模型中应用单调注意力来完成实时翻译。进一步地,Ma 等人^[40]提出了 MMA(Monotonic Multihead Attention),通过将单调注意力应用于 Transformer 架构,从而显著提升了实时翻译性能。

尽管单调注意力在实时翻译中取得了一定成效,但其复杂的期望训练过程仍然面临训练不稳定与计算复杂度较高的问题。为了解决这些问题,Zhang 等人^[41]为 MMA 提出了对偶路径框架(Dual-Path),通过引入不同翻译方向之间的对偶约束,增强了实时翻译策略的稳定性和质量。同时,Zaidi 等人^[42]引入了决策正则项,以提高单调注意力的训练稳定性。为进一步提升稳定性,Zhang 等人^[43]提出了 GMA(Gaussian Multihead Attention),通过为交叉注意力机制添加高斯先验,进一步优化了单调注意力的表现。针对某些语言的语序差异,Han 等人^[44]在单调注意力中引入了局部重排序机制。此外,Indurthi 等人^[45]在生成伯努利选择概率时结合了语言模型,从而进行更高质量的建模。针对训练效率,Ma 等人^[46]提出了高效 MMA,以提高 MMA 的训练效率。除了这些改进,一些新的基于注意力机制的架构也被提出用于实时翻译。Zhang 等人^[47]提出了 HMT(Hidden Markov Model),该方法将翻译的开始时机与生成的目标标记分别视为隐状态和观测状态,并利用隐马尔科夫模型进行建模。通过最大化观测状态的边际似然,HMT 能够自适应地学习最优的隐状态,从而确定翻译的开始时机。Ma 等人^[48]提出了 NAST,该方法通过非自回归架构建模源序列与目标序列的对齐关系,并学习最优的实时翻译策略。

基于 Transducer 的自适应策略 除了基于注意力机制的自适应策略之外,另一部分研究则着眼于通过 Transducer 架构实现实时翻译。如图 4(d)所示,Transducer 模型(如 RNN-T^[49]、T-T^[50])由编码器、预测器(predictor)和协同网络(joint network)组成。在协同网络中,空标记〈blank〉被添

加至输出词表。当生成<blank>时,模型继续等待源标记;而一旦生成非<blank>的目标标记,模型便输出该目标标记。训练过程中,Transducer 通过动态规划考虑所有可能包含<blank>的序列,从而实现最优的实时翻译策略。

在此框架下,Liu 等人^[51]提出了 CAAT,该方法移除了传统 Transducer 架构中的强单调约束,从而适应实时翻译任务的需求。借鉴 Transducer 的思路,Miao 等人^[52]提出了一种生成式框架 GSiMT,采用隐变量模型来建模读操作与写操作,并通过 Transducer 损失函数进行联合优化。最近,为了兼顾注意力机制和 Transducer 架构,Tang 等人^[53]提出了 TAED,该模型结合了 Transducer 在实时生成方面的优势与编码器-解码器在序列建模中的优点,进而提升了实时翻译性能。

自适应策略能够根据上下文动态决策何时开始翻译,从而显著提升实时翻译性能。在基于代理器、基于特殊标记、基于注意力机制和基于 Transducer 4 类方法中,基于注意力机制的自适应策略因其能够联合训练翻译模型与实时策略,且训练复杂度相对 Transducer 较低,已成为当前文本到文本实时翻译中最为主流的策略。

2.1.3 训练方式

实时策略使得模型能够在合适的时机开始翻译,如何提升模型在面对不完整源序列时的翻译能力,仍然是实时翻译面临的重大挑战。现有研究主要通过改进实时翻译模型的训练方式来提高其翻译效果。

引入未来信息 在训练过程中,直接使用不完整的源序列进行翻译,不仅具有较高的挑战性,还容易引发错误现象。大量研究发现:在训练阶段引入未来信息能够改善实时翻译模型的翻译效果^[2,54-56]。例如,一系列研究通过在训练中引入离线模型,利用未来信息对实时翻译模型进行词级蒸馏^[54]、序列级蒸馏^[55]以及表示层面蒸馏^[56],从而显著提升了训练效果。此外,部分研究进一步提出,在训练过程中通过语言模型任务或者预测任务显式提升实时翻译模型预测未来信息能力^[1,57-61],从而使得模型能够在推理过程中,从不完整的源序列中预测部分未来信息,进而提高实时翻译的质量。

减轻源信息不完整导致的训练错误 Zhang 等

人^[62]和 Zhong 等人^[63]指出,源信息的不完整可能导致编码器-解码器架构的翻译模型在注意力机制和生成过程等方面出现异常现象^[64]。针对注意力机制的异常,Zhang 等人^[62]提出了长度感知框架,通过掩码预测技术改善实时翻译模型的训练过程。针对生成目标词出现的异常,Liu 等人^[65]面向实时翻译模型提出了词级别的训练损失,旨在减少异常数据对训练效果的负面影响。另外,为了增强模型对不完整输入的鲁棒性,Raffel 等人^[66]提出了可移位上下文训练方法,该方法使得实时翻译模型能够在训练阶段学习到更多样的源信息情况,从而提高其在面对不完整信息时的翻译能力。

2.1.4 数据增强

除了实时策略和训练方法之外,部分研究从数据层面入手,旨在提升实时翻译模型的表现。其中,应用最为广泛的策略是将训练数据调整为更符合实时翻译场景特点的风格,从而增强模型在实时环境中的翻译能力。大多数研究集中在源序列与目标序列之间的对齐单调性上,致力于将目标序列改写为与源序列更加单调的形式,以适应实时翻译的要求^[67-70]。除单调性外,还有部分研究关注数据的整体风格,提出通过改写数据,加入省略、调序等特点,进一步提升模型在实时翻译任务中的表现^[71-72]。此外,针对一些特定语言,如中文^[73]和日文^[74],部分研究构造了针对性的数据集,旨在解决这些语言在实时翻译中的特殊需求。

2.1.5 优缺点分析

本节介绍了文本到文本实时翻译方法。固定策略,如 Wait- k 和其改进版本,提供了简单且易于实现的解决方案。其优势在于策略明确、计算开销低,其主要缺点是无法根据上下文调整策略。自适应策略,通过动态调整翻译时机,能够更好地应对源语言信息不完整的情况,从而提升翻译质量。而针对自适应方法存在的问题,现有方法从训练方式和数据增强的角度进行有针对性的改进,进一步提升文本到文本的实时翻译性能。

2.2 语音到文本实时翻译方法

在实际应用中,文本到文本实时翻译通常需要与语音识别模块级联。然而,这种级联可能导致误差累积和延迟增加等问题。因此,端到端语音到文本实时翻译被提出,旨在直接将源语言语音输入实

时翻译为目标语言文本。由于语音的连续性和多样性,所以语音到文本实时翻译在策略设计和翻译能力方面面临比文本到文本实时翻译更大的挑战。语音到文本实时翻译的研究主要集中在 3 个方向:分段方法、端到端策略、离线模型适配。分段方法通过预先将语音输入划分为词级别的片段,并在这些片段上直接应用第二节中介绍的文本实时翻译方法。端到端策略则免去了预先分段的过程,直接在源语音和目标文本之间构建实时翻译策略。离线模型适配方法则依赖于离线语音翻译模型,并结合推理策略实现语音到文本的实时翻译。本小节将对这三种方法进行详细的探讨和分析。

2.2.1 分段方法

语音到文本实时翻译的核心挑战在于语音序列的连续性及其远长于文本序列的特性。这使得在连续的语音流中找到合适的时机开始翻译变得更加复杂。为应对这一挑战,分段方法被提出将输入的语音进行实时分段,并将每个语音片段视为一个整体,进而在其上应用实时翻译策略。分段方法可分为固定分段和动态分段两类。

固定分段 固定分段方法通过将语音输入按固定长度划分为若干语音片段,例如,每 280ms 划分一个语音片段^[75]。在此基础上,可以在每个语音片段上应用 Wait- k 策略^[2]、MMA^[40] 等实时翻译策略。尽管固定分段方法实现简单,但由于其固定长度的特性可能破坏语音的自然结构,从而影响翻译的准确性。

动态分段 相比之下,动态分段方法根据语音输入的实际内容进行动态决策,以在更精确的位置对语音进行分段。早期的研究采用启发式方法来确定分段位置。例如,一些研究通过检测语音中的单词边界^[3,76] 来确定分段位置,而另一些研究则利用带有空标记<blank>的连接时序分类方法,识别语音中的停顿^[77-78],作为分段依据。除此之外,Zhang 等人^[79] 提出了 MU-ST,利用外部分段数据训练分段模型,从而在实时翻译中实现语音分段。在上述方法中,分段模块往往和翻译模型相独立。为了更有效地联合训练分段模块与翻译模型,Zhang 等人^[80] 提出了可微分割 DiSeg 方法,将语音切分过程转化为可微形式,并通过期望训练与翻译模型联合训练,学习出最适合翻译任务的分段策略。上述这些动态分段方法通常在分段之后执行 Wait- k 策略,以实现

语音到文本的实时翻译。

2.2.2 端到端策略

端到端策略通过直接在语音帧上建模实时翻译,避免了预先进行分段的步骤。大多数端到端策略依赖于整合并激发(Integrate-and-Fire)算法,该算法通过预测每个语音帧的信息量,并在累计的信息量超过设定阈值时启动翻译^[81-82]。该方法的优势在于其直接在语音帧层面进行实时翻译,具有较好的实时性。然而,基于整合并激发的策略往往忽略了翻译过程,仅关注输入的语音信息,从而难以获得对翻译最优的实时策略。为了解决这一问题,Zhang 等人^[83] 提出了 ITST,将翻译过程视为信息从源语音向目标文本的运输过程。只有当目标文本接收到的累计信息量超过某一阈值时,翻译才会被触发。与整合并激发算法不同,ITST 更加注重翻译过程中的信息传递,能够在实际翻译中提供更加精准的实时翻译策略。除了在对语音帧进行决策的基础上,续等待还是开始翻译,Zhang 等人^[84] 提出 Seg2Seg 框架,将相邻的若干语音帧视作一个整体片段进行决策,从而能够考虑更多的上下文信息,确保语音的局部完整性。

此外,一些研究通过翻译过程中的交叉注意力机制来指导实时翻译策略^[85-86],通过监控注意力的累计变化,确保在注意力达到设定阈值时启动翻译过程。还有一些研究借鉴了文本到文本实时翻译的成功经验,采用 Transducer 架构来建模语音到文本实时翻译中的实时策略^[51,53,87-88],从而同时优化实时策略和翻译过程。

最近,一些研究认为实际场景中实时翻译的输入则是连续的、未分段的文本/音频,因此开始关注在未分段的流式语音上进行翻译(Streaming Translation)。Iranzo Sanchez 等人^[89] 将最先进的实时文本机器翻译系统扩展到未分段的流式输入设置下,取得了良好的结果。Papi 等人^[90] 提出了 StreamAtt,其可以处理连续和无限制语音流输入,并根据翻译内容动态地对历史上下文进行选择。该方向的研究目前仍处于前沿探索阶段。

2.2.3 离线模型适配

在语音到文本实时翻译领域,除了从头开始构建翻译模型外,部分研究开始探索如何将训练好的离线语音翻译模型应用于实时翻译^[86,91]。然而,由于离线模型在训练过程中通常能够访问完整

的语音输入,训练-测试的不匹配在实际应用中可能导致翻译质量的不稳定性。为了应对这一问题,Liu 等人^[92]和 Dugan 等人^[93]提出了一种控制策略,避免离线模型在实时翻译中直接输出译文,而是仅选择性地输出已经足够稳定的译文。这种方式可以有效减少翻译过程中的不确定性,从而提高翻译的稳定性和准确性。另外,另一部分研究则通过改进束搜索来增强离线模型的实时适应能力^[94-96],通过选择包含公共前缀等置信度较高的译文,从而提高实时翻译的准确性。

与从头训练模型相比,离线模型适配方法的主要优势在于其能够快速应用现有的离线翻译系统,具有较低的开发成本和较短的部署周期。然而,这一方法仍面临训练-测试不匹配的挑战。因此,如何进一步提高离线模型的适配性,仍然是这一方向中亟待解决的问题。

2.2.4 优缺点比较

本节介绍了语音到文本实时翻译方法。分段方法的优势在于其实现简单且可以借助之前的实时策略。端到端策略避免了语音预分段过程,直接在语音帧层面进行翻译,能够取得更精确的实时策略。另外,一些离线模型适配研究旨在降低开发成本,缩短部署周期,但由于训练与测试数据的不匹配,这种方法在实际应用中仍然存在改进空间。整体来看,目前的端到端策略具有较高的实时翻译性能,因而成为主流的语音到文本实时翻译方法。

2.3 语音到语音实时翻译方法

随着实时场景应用的不断增多,诸如同声传译、在线直播等任务对实时翻译模型提出了更高的要求,希望模型能够直接生成目标语言语音,从而为用户提供更加全面的低延时沟通体验。然而,语音信号的多样性使得语音到语音实时翻译成为当前实时翻译领域最具挑战性和前沿性的任务之一。现有研究主要分为 3 类:模型架构、端侧模型以及音色保留。每一类研究在目标和应用上都有所不同,模型架构主要聚焦于设计能够支持高质量语音到语音实时翻译的框架,端侧模型则致力于在计算资源有限的环境中实现高效且高质量的翻译,而音色保留则注重在翻译过程中保持说话人的音色特征。本

小节将对上述三类研究方向进行详细的介绍与讨论。

2.3.1 模型架构

在语音到语音实时翻译的研究中,现有方法通常依赖于 Hubert^[47]等模型从源语音中提取 unit 序列,这些离散化的 unit 序列可以表示语音内容。具体来说,语音到语音实时翻译的数据源语言语音 S , 和目标语言语音的 unit 序列 U 。基于此,语音到语音翻译模型的目标转换为生成目标语言的 unit 序列 U ,之后即可使用声码器(Vocoder)基于 Unit 序列 U 合成目标语言的语音。现有的语音到语音实时翻译架构主要包括一阶段架构和两阶段架构。

一阶段架构 如图 5(a)所示,一阶段架构通过编码器-解码器结构直接将源语言语音 S 映射为目标语言语音 U 。这一架构与文本到文本实时翻译中的编码器-解码器架构相似,只是输入数据的类型不同。因此,某些原本用于文本翻译的实时策略可以直接迁移至语音到语音翻译中。例如,Ma 等人^[97]基于 2.1 节中的 NAST^[48]提出了 NAST-S2x,采用非自回归解码器处理实时翻译。Zhao 等人^[98]RNN-T 进行语音到语音实时翻译,取得了良好的效果。然而,由于一阶段架构的建模能力稍弱,此类方法的实时翻译性能也受到局限。

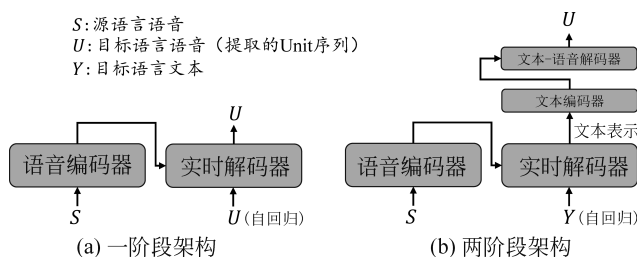


图 5 语音到语音实时翻译一阶段和两阶段架构图

两阶段架构 如图 5(b)所示,两阶段架构最早由 UnitY 框架提出^[99],并已成为语音到语音实时翻译中最为常用的架构。在这一架构中,第一阶段将源语音映射为目标语言的文本表示,第二阶段则通过目标文本表示合成目标语音。两阶段架构可以端到端训练,增强了模型的建模能力。在 UnitY 架构的基础上,Zhang 等人^[4]提出了 StreamSpeech,该方法采用多任务学习框架联合学习语音到语音翻译

与实时策略。StreamSpeech 引入了若干 CTC 对齐机制,用于指导翻译何时开始,并通过实时语音识别、实时语音翻译等辅助任务来训练这些对齐机制。类似地,Seamless-Communication^[100]提出 Seamless 模型,采用高效 MMA^[46]作为实时策略,并在大规模数据集上训练了一个具有 10 亿参数的语音到语音实时翻译模型。目前,Seamless 被认为是性能最优的语音到语音实时翻译模型。

2.3.2 端侧模型

端侧模型旨在在计算资源和内存空间有限的环境里实现高效且高质量的语音到语音实时翻译。此类模型的研究重点在于如何在边缘设备或资源受限的场景中进行推理,既能确保计算开销较小,又能维持翻译的质量。目前针对端侧语音到语音实时翻译模型的研究仍处于前沿阶段,但已有一些探索取得了不错的进展。Agranovich 等人^[101]提出 SimulTron,将 Translatotron 2^[102]适配为实时翻译模型,并通过将其关键组件替换为轻量级的 LSTM 网络,成功地展示了在移动设备上进行语音到语音实时翻译的潜力。

2.3.3 音色保留

与语音到文本实时翻译相比,语音到语音实时翻译为音色保留提供了更大的可能性。音色保留的目标是确保翻译后的语音能够保持源语音的音质、语气、语调等特征,从而使翻译结果更加自然且贴近源语言的表达方式。这一任务在语音到语音实时翻译中尤为重要,因为音色不仅仅影响语言的传递,也与说话人的身份和情感传达紧密相关。例如,Le 等人^[103]提出的 TransVIP 通过引入两个分离的编码器,在翻译过程中专门保留源语音中的说话者音色特征及语音节奏性,确保翻译质量与说话者音色的良好保留。类似地,Jia 等人^[102]提出了 Translatotron 2,在模型中增设了一个单独的声学合成器,用以在翻译过程中保持每个说话者的音色特点。这些方法为语音到语音翻译中的音色保留提供了有效的解决方案。Labiausse 等人^[104]提出的 Hibiki 采用了一种弱监督学习方法,通过分析生成的困惑度来优化每个词的翻译延迟,从而在实时翻译中适应不同的语境和说话节奏。

2.3.4 优缺点比较

现有的一阶段架构虽然结构简洁,易于实现,但

由于其建模能力相对较弱,难以在复杂场景中保证高质量的翻译。相比之下,两阶段架构通过将源语音转化为目标语言文本,再进行语音合成,能够有效提升翻译的准确性。端侧模型则在资源有限的环境下提供了一种轻量级的解决方案,能够在移动设备等边缘设备上实现高效翻译,但其性能通常难以与大模型相媲美。音色保留技术为语音到语音翻译带来了更自然的输出,通过保持源语音的音质、语气和语调,进一步增强了翻译结果的真实感和情感传达。总体而言,语音到语音实时翻译相较于文本到文本翻译和语音到文本翻译,能够为用户提供更全面的低延时沟通体验。此外,端到端架构为音色保留提供了更大的可能性,进一步提高了翻译的自然度和准确性。尽管目前语音到语音实时翻译已取得一定进展,但相关研究仍处于起步阶段,仍然是实时翻译领域中最具挑战性和前沿性的研究方向。

2.4 基于大模型的实时翻译方法

随着大模型^[105](Large Language Model, LLM)的迅猛发展,近年来的研究逐渐转向探索如何利用大模型实现实时翻译。然而,大模型通常采用指令-回复的交互模式,这种一问一答的方式与实时翻译要求输入与输出同步进行的特点存在一定的矛盾。因此,如何将大模型有效适配于实时翻译任务,并发挥其强大的生成能力,成了相关研究的核心问题。基于大模型的实时翻译方法大体可分为三类:上下文学习、指令微调和智能体方法。本小节将对这三类研究方向进行详细介绍与讨论。

2.4.1 上下文学习

利用上下文学习方法完成实时翻译任务的关键在于如何引导大模型遵循实时翻译任务指令。由于大模型通常是在完整句子或段落上进行训练,直接让其翻译不完整的源序列可能会导致翻译结果的不稳定或产生幻觉。为此,Koshkin 等人^[5]在提示(prompt)中提供当前部分输入、之前部分翻译及附加背景信息,引导 LLM 生成翻译,如图 6(a)所示。随着翻译过程的推进,输入的提示会不断更新,包含新的源语句和翻译结果,从而有助于提高翻译的稳定性和一致性。尽管此方法在一定程度上能够改善翻译质量,但仍然存在不稳定的潜在风险。Wang

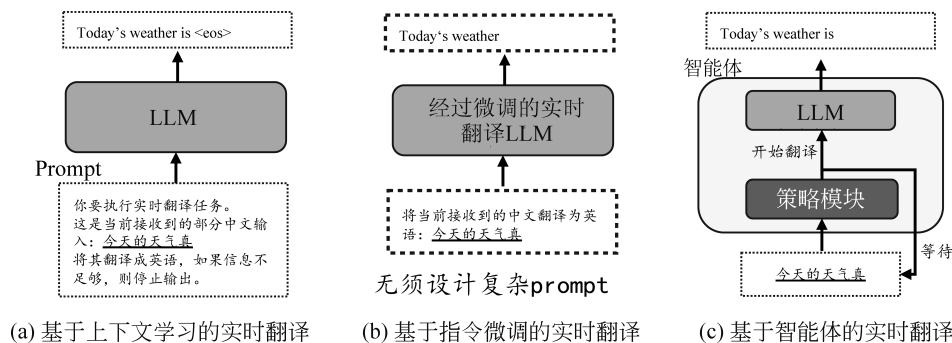


图6 基于大模型的实时翻译示意图

等人^[106]提出最长公共前缀算法(RALCP),旨在从LLM的输出中提取最有可能的翻译结果,以提高实时翻译的稳定性。上述上下文学习的方法通常依赖单轮问答,在每一次翻译中需要重新编码提示,未能充分利用LLM的缓存机制,导致翻译效率较低。为此,Wang等人^[107]提出了多轮对话提示结构,避免重复计算提示,从而显著提升了翻译效率。基于上下文学习的方法往往只能依赖预先设定的固定策略,限制了实时翻译性能。

2.4.2 指令微调

指令微调作为一种直接适配LLM于实时翻译的方法,面临的主要挑战是如何突破大模型一问一答交互的范式,以满足实时翻译对低延迟的要求。如图6(b)所示,现有方法在指令微调过程中对LLM的训练数据、训练流程或架构进行适当调整,以适配实时翻译。Agostinelli等人^[108]采用Wait- k 策略离线构建包含不完整源序列和目标序列的实时翻译数据集,对LLM进行有监督微调,从而让模型能够在源信息不完全的情况下进行翻译。然而,这种数据驱动的方法往往只能实现固定的翻译策略,难以灵活应对动态变化的实时翻译场景。为了解决这一问题,Raffel等人^[109]提出了SimulMask微调范式,通过在训练过程中掩码关注区域来模拟实时翻译,并利用掩码技术实现不同的翻译策略。此外,部分研究从架构层面进行了改进,Koshkin等人^[110]提出了TransLLaMA,在LLM的词表中引入了等待标记<WAIT>,使得模型能够在源信息不足时通过生成<WAIT>等待更多信息。为了进一步优化实时翻译性能,Yu等人^[111]提出了SimulPL框架,从翻译质量、单调性、关键信息、简洁性和延迟等多个维度对实时翻译中的人类偏好进行细分,并基于这

些偏好通过直接偏好优化算法(DPO)对LLM进行微调,进一步提升了实时翻译质量。尽管指令微调在提高实时翻译性能方面取得了一定进展,目前方法仍然较为简单直接,未来仍有大量空间来探索如何在翻译质量和延迟之间取得更优的权衡。

2.4.3 智能体

实时翻译任务的复杂性较高,单纯依赖LLM往往难以完成。因此,部分研究开始探索智能体方法,通过将策略模块与LLM结合,更灵活地控制翻译过程。如图6(c)所示,Guo等人^[112]提出了Agent-SiMT,利用策略模块与LLM共同完成实时翻译。策略模块使用自适应策略HMT^[47]基于当前上下文与翻译历史判断是否继续等待或开始翻译。通过解耦策略和翻译过程,Agent-SiMT能够有效地利用LLM的生成能力和策略模型的决策能力。为进一步提升翻译系统的拟人化水平,Cheng等人^[113]提出了CLASI,借鉴了专业人类口译员的翻译风格,设计了一种拟人化的实时翻译方法。CLASI引入了一个多模态检索模块,能够根据音频输入、历史上下文和相关检索信息增强翻译的容错性,尤其在面对领域特定术语时表现优异。智能体方法通过结合生成能力与策略决策显著提升了实时翻译系统在复杂场景中的表现,尤其在翻译质量和拟人化方面取得了显著的性能提升。

2.4.4 优缺点比较

基于上下文学习的方法虽然简单,但是往往只能依赖一种固定策略进行实时翻译,导致实时翻译质量难以达到最佳。而指令微调方法通过对LLM进行微调,使其具备实时翻译能力,其优势在于能够根据实际需求调整训练数据和策略以进行特定的优

化,同时能够建模人类在实时翻译中的偏好。基于智能体的方法通过策略模块与 LLM 相结合,能够更灵活地控制翻译过程,有效提升了系统在复杂场景中的表现。目前,基于大模型的实时翻译方法仍处于探索阶段。未来研究可以从实时策略、LLM 效率、人类偏好等方面入手,进一步提升实时翻译的性能。

2.5 现有实时翻译方法性能对比

本文评估了大部分主流实时翻译方法的性能,

旨在为研究人员提供横向对比。对于文本到文本实时翻译方法和基于大模型的实时翻译方法,测试集采用 WMT15 德语到英语翻译^[114],延时指标单位为词数。对于语音到文本实时翻译方法,测试集采用 MuST-C 英语到德语翻译^[115],延时指标单位为毫秒。对于语音到语音实时翻译方法,测试集采用 CVSS-C 法语到英语翻译^[116],延时指标单位为毫秒。各个模型的评估设置均采用其原文的标准设置,并选取低延时和高延时两个延时水平。评估结果如表 1 所示。

表 1 实时翻译方法的性能对比

类别	方法	拟解决挑战	是否支持其他实时任务	实时翻译性能			
				低延时		高延时	
				延时 ↓ (AL)	翻译质量 ↑ (BLEU)	延时 ↓ (AL)	翻译质量 ↑ (BLEU)
文本到文本 实时翻译 (测试集: WMT15 德语- 英语;延时单位为词数)	Offline ^[9]	—	—	—	—	24.9	31.6
	Wait- <i>k</i> ^[2]	固定策略	✓	3.9	26.9	11.5	30.8
	Multipath Wait- <i>k</i> ^[18]	固定策略	×	3.7	26.9	11.5	30.3
	Adaptive Wait- <i>k</i> ^[18]	固定策略	×	4.1	27.4	11.6	30.8
	MoE Wait- <i>k</i> ^[19]	固定策略	×	3.9	28.3	11.5	31.1
	PED ^[14]	训练方法	×	3.3	27.8	12.2	29.9
	Wait-info ^[17]	自适应策略	×	3.8	28.4	8.4	28.8
	MMA ^[40]	自适应策略	✓	4.9	27.5	13.2	28.7
	Dual-Path ^[41]	训练方法	×	3.2	27.0	13.5	30.1
	LAF ^[62]	训练方法	×	4.6	28.7	12.9	29.8
	GMA ^[43]	自适应策略	×	3.9	28.1	14.0	30.3
	GSiMT ^[52]	自适应策略	×	3.6	28.8	6.3	29.6
	ITST ^[83]	自适应策略	✓	3.8	29.6	10.2	31.9
	ITST+KD ^[69]	数据增强	×	3.5	30.1	13.2	32.2
	HMT ^[47]	自适应策略	×	3.5	29.3	11.8	31.7
	DaP-SiMT ^[34]	训练方式	×	3.3	29.1	12.5	31.8
	Tailor ^[72]	训练方法	×	3.4	29.9	8.2	31.8
	CAAT ^[51]	自适应策略	✓	3.4	27.5	7.6	30.3
	Seg2Seg ^[84]	自适应策略	✓	3.8	29.3	11.2	32.2
	NAST ^[48]	自适应策略	✓	3.0	30.5	8.4	31.9
	Glance-HMT ^[56]	训练方法	×	3.0	28.5	6.5	30.3
	SM2 ^[26]	自适应策略	×	3.7	30.2	12.3	31.9
	PsFuture ^[60]	训练方法	×	3.7	29.6	12.3	32.5

续表

类别	方法	拟解决挑战	是否支持其他实时任务	实时翻译性能			
				低延时		高延时	
				延时↓ (AL)	翻译质量↑ (BLEU)	延时↓ (AL)	翻译质量↑ (BLEU)
语音到文本 实时翻译 (测试集: MuST-C 英语-德语; 延时单位为毫秒)	Offline ^[75]	—	—	—	—	5 654	24.3
	Pre-decision+Wait-k ^[75]	分段方法	×	1 385	18.1	3 540	20.3
	Wait-k-Stride-n ^[78]	分段方法	×	2 319	19.7	3 589	20.3
	Pre-decision+MMA ^[75]	分段方法	×	1 530	13.1	4 225	16.1
	MMA-CMDR ^[42]	分段方法	×	1 504	16.1	2 614	21.0
	RealTrans ^[78]	分段方法	×	1 355	16.5	3 106	20.4
	SH ^[77]	分段方法	×	1 320	18.0	2 490	20.0
	MoSST ^[82]	端到端策略	×	1 734	16.3	2 742	20.0
	ITST ^[83]	端到端策略	✓	1 588	18.5	3 501	22.3
	MU-ST ^[79]	分段方法	×	1 211	19.3	3 452	22.1
	DiSeg ^[80]	分段方法	×	1 514	20.7	3 464	23.1
	DiSeg + MGCM ^[64]	分段方法	×	1 492	19.4	3 310	23.2
	EdAtt ^[86]	离线模型适配	×	1 287	15.9	3 736	19.9
	AlignAtt ^[85]	端到端策略	×	1 900	19.3	3 354	27.8
	AlignAtt+CMF ^[96]	端到端策略	×	1 930	22.7	3 410	27.4
	CAAT ^[51]	端到端策略	✓	1 443	22.2	2 138	22.6
	Seg2Seg ^[84]	端到端策略	✓	1 602	21.7	3 125	23.6
	TAED ^[53]	端到端策略	✓	1 263	21.6	1 903	23.5
语音到语音 实时翻译 (测试集: CVSS-C 法语-英语; 延时单位为毫秒)	Offline (Translatotron 2 ^[102])	—	—	—	—	5 683	25.5
	Offline (UnitY ^[99])	—	—	—	—	5 683	26.9
	UnitY+Wait-k ^[4]	模型架构	×	1 286	11.9	3 994	19.8
	StreamSpeech ^[4]	模型架构	✓	1 269	22.9	3 812	25.9
	NAST-s2x ^[97]	模型架构	✓	1 533	25.7	3 330	25.7
	Seamless ^[100]	模型架构	✓	2 800	33.8	—	—
	HiBiKi ^[104]	模型架构	×	—	—	5 100	26.7
基于大模型的实时翻译 (测试集: WMT15 德语-英语; 延时单位为词数)	Offline ^[9]	—	—	—	—	24.9	31.6
	RALCP ^[106]	上下文学习	×	3.8	19.4	12.1	28.3
	RALCP+TAF ^[61]	指令微调	×	3.8	27.9	12.3	29.4
	ConverPrompt ^[107]	上下文学习	×	3.1	29.4	7.8	32.3
	Agent-SiMT ^[112]	智能体	×	4.2	30.6	9.9	33.4
	SimulPL ^[111]	指令微调	×	4.1	49.2	12.2	52.3

梳理实时翻译未来发展的三个关键方向。

3 展望与分析

在人工智能和自然语言处理技术迅猛发展的背景下,实时翻译系统正朝着更高的翻译准确性、更低的响应延迟以及更拟人化的方向不断发展。本节将

3.1 建立更全面的评估体系

现有研究中,实时翻译的主要目标一直聚焦于在低延迟下实现高质量翻译,但随着技术的不断进步,现有的评估指标已逐渐无法满足日益复杂的实

时翻译场景。传统的评估方法主要关注翻译质量和延迟,然而,随着应用场景的多样化,评估标准的单一性和粗糙性无法全面反映实时翻译技术的实际表现。例如,简单的延迟和质量评分难以衡量系统在多模态、语境变化或用户交互等复杂情况下的鲁棒性和适应能力。因此,随着技术的发展,评估指标的多样性和细致性将变得更加重要,且对评估方法的精确化和全面化提出了更高的要求。

近年来,部分研究开始提出更加精细化的实时翻译评价指标^[117-119],这些指标不仅考量翻译结果的准确度和延时,还包括上下文适应性、翻译风格等方面的表现。此外,一些研究也开始深入分析实时翻译中的错误类型^[120],例如,漏译、错译、翻译偏差等,以期更加全面地评估模型在不同应用场景中的表现。然而,实时翻译的评估复杂性要求研究人员未来探索更加精细化的评价方法,如基于大模型的评估标准、考虑不同偏好的评估框架等。未来,随着更加全面的实时翻译评估体系的建立,相关研究也将推动实时翻译技术的持续优化与发展,并为该领域的未来研究方向指明路径。

3.2 构建拟人化的实时翻译

尽管现有的实时翻译技术在翻译质量和延迟方面已经取得了显著的进展,但在同声传译、跨国旅游等具体应用场景中仍有提升空间,特别是实时翻译的拟人化表现^[113]。拟人化实时翻译的核心目标是模拟人类译员在翻译过程中的特征,从而提升翻译结果的情感表达和语言自然度,使得机器生成的语音听起来更具人类特色。

具体来说,拟人化翻译不仅仅是将文字信息转化为语音,更是要在音调、语速、语气等方面进行细致调整,使得合成的语音能够更好地传达说话人的情感和语气。例如,模拟人类的副语言现象,如呼吸、叹气、语速变化、情感波动等,可以使得翻译后的语音听起来更加自然、亲切,避免了目前很多机器翻译语音显得生硬、机械的问题。此外,一些研究^[102]已经开始关注如何在翻译过程中保留音色特征,使翻译语音不仅在内容上准确,而且在节奏、情感和语气上与原始音频更为贴合,从而使得机器生成的语音更富有表现力,增强听者的代入感和接受度。未

来,随着在这一方向的深入探索,实时翻译将能够提供更为丰富和个性化的声音表现形式,不仅使得翻译更为精确,也将在跨文化沟通中为用户带来更为人性化的互动体验。

3.3 拓展至实时交互大模型

随着大模型技术的迅猛发展,实时交互大模型的构建已经成为人工智能领域中的一个重要研究方向。在这一过程中,实时翻译技术起到了至关重要的推动作用。实时翻译中的实时策略不仅具备坚实的理论基础,还具有良好的可迁移性,能够有效地支持实时交互大模型的实现。具体来说,实时翻译系统的核心优势在于其能够处理连续输入、快速响应,并实时输出翻译结果,这一特性与大模型在处理复杂、动态输入时的需求高度契合。

目前,已有研究将实时翻译中的相关技术迁移至大模型体系中^[121-122],构建出能够快速响应的实时交互系统,并取得了良好的效果。展望未来,随着实时翻译技术与大模型技术的不断演进,利用现有的实时翻译技术指导构建实时交互大模型将成为未来研究的一个重要热点。这一趋势有望推动人机交互模式的转变,实现更加自然和高效的实时沟通。

4 总结

本文系统地梳理了近年来实时翻译领域的研究进展,并深入分析了各类方法的优缺点。同时,本文展望了实时翻译的未来发展方向,提出了潜在的技术突破与发展趋势。最后,本文对相关文献进行了分类整理,按方法与发表时间进行归类,以期为研究人员提供全面的参考和研究方向指引,推动实时翻译技术的进一步发展与应用。

参考文献

- [1] GRISSOM II A, HE H, BOYD GRABER J, et al. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation [C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics,

- 2014: 1342-1352.
- [2] MA M, HUANG L, XIONG H, et al. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework [C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Italy: Association for Computational Linguistics, 2019: 3025-3036.
- [3] REN Y, LIU J, TAN X, et al. SimulSpeech: End-to-end simultaneous speech to text translation [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3787-3796.
- [4] ZHANG S, FANG Q, GUO S, et al. StreamSpeech: Simultaneous speech-to-speech translation with multi-task learning [C/OL]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand: Association for Computational Linguistics, 2024: 8964-8986.
- [5] KOSHKIN R, SUDOH K, NAKAMURA S. LLMs are zero-shot context-aware simultaneous translators [C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA: Association for Computational Linguistics, 2024: 1192-1207.
- [6] 冯洋. 神经机器翻译前沿综述[J]. 中文信息学报, 2020, 34(7): 1-18.
- [7] 李亚超, 熊德意, 张民. 神经机器翻译综述[J]. 计算机学报, 2018, 41(12): 2734-2755.
- [8] XU C, YE R, DONG Q, et al. Recent advances in direct speech-to-text translation [C/OL]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence, Organization, 2023: 6796-6804.
- [9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017: 5998-6008.
- [10] PAPI S, GAIDO M, NEGRI M, et al. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation [C/OL]//Proceedings of the 3rd Workshop on Automatic Simultaneous Translation, 2022: 12-17.
- [11] CHO K, ESIPOVA M. Can neural machine translation do simultaneous translation? [A/OL]. arXiv: 1606.02012, 2016.
- [12] DALVI F, DURRANI N, SAJJAD H, et al. Incremental decoding and training methods for simultaneous translation in neural machine translation [C/OL]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA: Association for Computational Linguistics, 2018: 493-499.
- [13] ZHENG B, LIU K, ZHENG R, et al. Simultaneous translation policies: From fixed to adaptive [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 2847-2853.
- [14] GUO S, ZHANG S, FENG Y. Turning fixed to adaptive: Integrating post-evaluation into simultaneous machine translation [C/OL]//Proceedings of the Association for Computational Linguistics Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 2264-2278.
- [15] ZHANG S, FENG Y. ICT's system for AutoSimTrans 2021: Robust char-level simultaneous translation [C/OL]//Proceedings of the 2nd Workshop on Automatic Simultaneous Translation, Online: Association for Computational Linguistics, 2021: 1-11.
- [16] KIM K, CHO H. Enhanced simultaneous machine translation with word-level policies [C/OL]//Proceedings of the Association for Computational Linguistics, Singapore: Association for Computational Linguistics, 2023: 15616-15634.
- [17] ZHANG S, GUO S, FENG Y. Wait-info policy: Balancing source and target at information level for simultaneous machine translation [C/OL]//Proceedings of the Association for Computational Linguistics, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 2249-2263.
- [18] ELBAYAD M, BESACIER L, VERBEEK J. Efficient Wait- k models for simultaneous machine translation [C/OL]//Proceedings of the Interspeech, 2020: 1461-1465.
- [19] ZHANG S, FENG Y. Universal simultaneous machine translation with mixture-of-experts Wait- k policy [C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 7306-7317.
- [20] SATIJA H, PINEAU J. Simultaneous machine translation using deep reinforcement learning [J/OL]. arXiv: 1610.00388, 2016.
- [21] GU J, NEUBIG G, CHO K, et al. Learning to translate

- in real-time with neural machine translation[C/OL]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 2017: 1053-1062.
- [22] IVE J, LI A M, MIAO Y, et al. Exploiting multimodal reinforcement learning for simultaneous machine translation[C/OL]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021: 3222-3233.
- [23] LEE Y, SHIN J, KIM Y. Simultaneous neural machine translation with a reinforced attention mechanism[J]. Etri Journal, 2021, 43(5): 775-786.
- [24] ZHENG B, ZHENG R, MA M, et al. Simpler and faster learning of adaptive policies for simultaneous translation[C/OL]// Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 1349-1354.
- [25] ARTHUR P, COHN T, HAFFARI G. Learning coupled policies for simultaneous machine translation using imitation learning[C/OL]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online, 2021: 2709-2719.
- [26] YU D, KANG X, LIU Y, et al. Self-modifying state modeling for simultaneous machine translation[C/OL]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 2024: 9781-9795.
- [27] CHEN Y, LI L, JIANG X, et al. How to do simultaneous translation better with consecutive neural machine translation? [J/OL]. arXiv: 1911.03154, 2019.
- [28] CHEN Y, LI L, JIANG X, et al. A general framework for adaptation of neural machine translation to simultaneous translation [C/OL]// Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Suzhou, China, 2020: 191-200.
- [29] ZHANG R, ZHANG C, HE Z, et al. Learning adaptive segmentation policy for simultaneous translation [C/OL]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 2280-2289.
- [30] XIONG H, ZHANG R, ZHANG C, et al. Dutongchuan: Context-aware translation model for simultaneous interpreting[A/OL]. arXiv: 1907.12984, 2019.
- [31] KANO Y, SUDOH K, NAKAMURA S. Simultaneous neural machine translation with constituent label prediction[J/OL]. arXiv: 2110.13480, 2021.
- [32] ALINEJAD A, SHAVARANI H S, SARKAR A. Translation-based supervision for policy generation in simultaneous neural machine translation [C/OL]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, 2021: 1734-1744.
- [33] GUO S, ZHANG S, FENG Y. Learning optimal policy for simultaneous machine translation via binary search [C/OL]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 2318-2333.
- [34] ZHAO L, FAN K, LUO W, et al. Adaptive policy with wait- k model for simultaneous translation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 4816-4832.
- [35] ZHENG B, ZHENG R, MA M, et al. Simultaneous translation with flexible policy via restricted imitation learning[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 5816-5822.
- [36] CHOUSA K, SUDOH K, NAKAMURA S. Simultaneous neural machine translation using connectionist temporal classification[A/OL]. arXiv: 1911.11933, 2019.
- [37] RAFFEL C, LUONG M T, LIU P J, et al. Online and linear-time attention by enforcing monotonic alignments [C/OL]// Proceedings of the 34th International Conference on Machine Learning. PMLR, 2017: 2837-2846.
- [38] CHIU R. Monotonic chunkwise attention [C]// Proceedings of the International Conference on Learning Representations, 2018.
- [39] ARIVAZHAGAN N, CHERRY C, MACHEREY W, et al. Monotonic infinite lookback attention for simultaneous machine translation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 1313-1323.
- [40] MA X, PINO J M, CROSS J, et al. Monotonic multihead attention[C]// Proceedings of the International Conference on Learning Representations, 2020.
- [41] ZHANG S, FENG Y. Modeling dual read/write

- paths for simultaneous machine translation [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 2022: 2461-2477.
- [42] ZAIDI M A, LEE B, KIM S, et al. Cross-modal decision regularization for simultaneous speech translation [C]//Proceedings of the Interspeech, 2022: 116-120.
- [43] ZHANG S, FENG Y. Gaussian multi-head attention for simultaneous machine translation [C]// Proceedings of the Association for Computational Linguistics, Dublin, Ireland, 2022: 3019-3030.
- [44] HAN H, AHN S, CHOI Y, et al. Monotonic simultaneous translation with chunk-wise reordering and refinement [C]//Proceedings of the 6th Conference on Machine Translation, Online, 2021: 1110-1123.
- [45] INDURTHI S R, ZAIDI M A, LEE B, et al. Language model augmented monotonic attention for simultaneous translation [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022: 38-45.
- [46] MA X, PINO J M, CROSS J, et al. Monotonic multihead attention [C]//Proceedings of the International Conference on Learning Representations, Online, 2020: 1-16.
- [47] ZHANG S, FENG Y. Hidden Markov transformer for simultaneous machine translation [C]// Proceedings of the 11th International Conference on Learning Representations, 2023.
- [48] MA Z, ZHANG S, GUO S, et al. Non-autoregressive streaming transformer for simultaneous translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore, 2023: 5177-5190.
- [49] GRAVES A. Sequence transduction with recurrent neural networks [J/OL]. CoRR, 2012, abs/1211.3711.
- [50] ZHANG Q, LU H, SAK H, et al. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss [C]//Proceedings of the ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 7829-7833.
- [51] LIU D, DU M, LI X, et al. Cross attention augmented transducer networks for simultaneous translation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021: 39-55.
- [52] MIAO Y, BLUNSOM P, SPECIA L. A generative framework for simultaneous machine translation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021: 6697-6706.
- [53] TANG Y, SUN A, INAGUMA H, et al. Hybrid transducer and attention based encoder-decoder modeling for speech-to-text tasks [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 2023: 12441-12455.
- [54] ZHANG S, FENG Y, LI L. Future-guided incremental transformer for simultaneous translation [J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14428-14436.
- [55] WU X, XIA Y, WU L, et al. Learning to use future information in simultaneous translation [J/OL]. arXiv, 2021.
- [56] GUO S, ZHANG S, FENG Y. Glancing future for simultaneous machine translation [C]//Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2024: 11386-11390.
- [57] ALINEJAD A, SIAHBANI M, SARKAR A. Prediction improves simultaneous neural machine translation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018: 3022-3027.
- [58] VOGLER N, STEWART C, NEUBIG G. Lost in interpretation: Predicting untranslated terminology in simultaneous interpretation [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2019: 109-118.
- [59] YIN A, ZHONG T, LI H, et al. Language model is a branch predictor for simultaneous machine translation [A/OL]. arXiv: 2312.14488, 2023.
- [60] ZHAO L, LI J, ZENG Z. PsFuture: A pseudofuture-based zero-shot adaptive policy for simultaneous machine translation [C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA, 2024: 1869-1881.
- [61] OUYANG S, HRINCHUK O, CHEN Z, et al. Anticipating future with large language model for simultaneous machine translation [A/OL]. arXiv: 2410.22499, 2024.
- [62] ZHANG S, FENG Y. Reducing position bias in

- simultaneous machine translation with length-aware framework[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 2022: 6775-6788.
- [63] ZHONG M, CHEN K, XUE Z, et al. On the hallucination in simultaneous machine translation [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 2024: 730-742.
- [64] LIU X, DU Y, WANG J, et al. A modular-based strategy for mitigating gradient conflicts in simultaneous speech translation[A/OL].arXiv: 2409.15911, 2024.
- [65] LIU M, ZHANG W, LI X, et al.Cbsimt: Mitigating hallucination in simultaneous machine translation with weighted prefix-to-prefix training [A/OL], 2023. arXiv: 2311.03672.
- [66] RAFFEL M, PENNEY D, CHEN L. Shiftable context: Addressing training-inference context mismatch in simultaneous speech translation[C]//Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023: 28519-28530.
- [67] CHEN J, ZHENG R, KITA A, et al. Improving simultaneous translation by incorporating pseudoreferences with fewer reorderings [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, 2021: 5857-5864.
- [68] XUN G, MA M, BIAN Y, et al. Data-driven adaptive simultaneous machine translation[A/OL].arXiv: 2204.12672, 2022.
- [69] DENG H, DING L, LIU X, et al. Improving simultaneous machine translation with mon-olingual data[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 12728-12736.
- [70] WANG S, WU J, FAN K, et al. Better simultaneous translation with monotonic knowledge distillation [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 2023: 2334-2349.
- [71] ZHAO J, ARTHUR P, HAFFARI G, et al. It is not as good as you think!: Evaluating simultaneous machine translation on interpretation data[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, 2021: 6707-6715.
- [72] GUO S, ZHANG S, FENG Y. Simultaneous machine translation with tailored reference[C]//Proceedings of the Association for Computational Linguistics, Singapore, 2023: 3070-3084.
- [73] ZHANG R, WANG X, ZHANG C, et al. BSTC: A large scale Chinese-English speech translation dataset [C]//Proceedings of the 2nd Workshop on Automatic Simultaneous Translation. Online: Association for Computational Linguistics, 2021: 28-35.
- [74] DOI K, SUDOH K, NAKAMURA S. Large-scale English-Japanese simultaneous interpretation corpus: construction and analyses with sentence-aligned data [C]//Proceedings of the 18th International Conference on Spoken Language Translation. Bangkok, Thailand (online), 2021: 226-235.
- [75] MA X, PINO J, KOEHN P. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation[C]//Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Suzhou, China, 2020: 582-587.
- [76] IRANZO SÁNCHEZ J, GIMÉNEZ PASTOR A, SILVESTRE CERDÀ J A, et al. Direct segmentation models for streaming speech translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Online, 2020: 2599-2611.
- [77] CHEN J, MA M, ZHENG R, et al. Direct simultaneous speech-to-text translation assisted by synchronized streaming ASR[C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP. Online, 2021: 4618-4624.
- [78] ZENG X, LI L, LIU Q. RealTranS: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer[C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP. Online, 2021: 2461-2474.
- [79] ZHANG R, HE Z, WU H, et al. Learning adaptive segmentation policy for end-to-end simultaneous translation [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 2022: 7862-7874.
- [80] ZHANG S, FENG Y. End-to-end simultaneous speech translation with differentiable segmentation [C]//Proceedings of the Association for Computational Linguistics: Toronto, Canada, 2023: 7659-7680.
- [81] CHANG C C, YI LEE H. Exploring continuous integrate-and-fire for adaptive simultaneous speech translation [C]//Proceedings of the Interspeech, 2022: 5175-5179.

- [82] DONG Q, ZHU Y, WANG M, et al. Learning when to translate for streaming speech [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 2022: 680-694.
- [83] ZHANG S, FENG Y. Information-transport-based policy for simultaneous translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 992-1013.
- [84] ZHANG S, FENG Y. Unified segment-to-segment framework for simultaneous sequence generation [C/OL]//Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, 2023: 45235-45258.
- [85] PAPI S, NEGRI M, TURCHI M. Attention as a guide for simultaneous speech translation [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 13340-13356.
- [86] PAPI S, TURCHI M, NEGRI M. Alignatt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation [C]//Proceedings of the Interspeech, 2023: 3974-3978.
- [87] WANG P, SUN E, XUE J, et al. Lamassu: A streaming language-agnostic multilingual speech recognition and translation model using neural transducers [C]//Proceedings of the Interspeech, 2023: 57-61.
- [88] DENG K, WOODLAND P. Label-synchronous neural transducer for E2E simultaneous speech translation [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024: 8235-8251.
- [89] IRANZO SANCHEZ J, CIVERA J, JUAN CÍSCAR A. From simultaneous to streaming machine translation by leveraging streaming history [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 2022: 6972-6985.
- [90] PAPI S, GAIDO M, NEGRI M, et al. StreamAtt: Direct streaming speech-to-text translation with attention-based audio history selection [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024: 3692-3707.
- [91] PAPI S, GAIDO M, NEGRI M, et al. Does simultaneous speech translation need simultaneous models? [C]//Proceedings of the Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022: 141-153.
- [92] LIU D, SPANAKIS G, NIEHUES J. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection [C]//Proceedings of the Interspeech, 2020: 3620-3624.
- [93] DUGAN L, WADHAWAN A, SPENCE K, et al. Learning when to speak: Latency and quality trade-offs for simultaneous speech-to-speech translation with offline models [C]//Proceedings of the Interspeech, 2023: 5265-5266.
- [94] ZHENG R, MA M, ZHENG B, et al. Speculative beam search for simultaneous translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 1395-1402.
- [95] POLÁK P, YAN B, WATANABE S, et al. Incremental blockwise beam search for simultaneous speech translation with controllable quality-latency tradeoff [C]//Proceedings of the Interspeech, 2023: 3979-3983.
- [96] TAN H, SAKTI S. Contrastive feedback mechanism for simultaneous speech translation [C]//Proceedings of the Interspeech, 2024: 852-856.
- [97] MA Z, FANG Q, ZHANG S, et al. A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024: 1557-1575.
- [98] ZHAO J, MORITZ N, LAKOMKIN E, et al. Textless streaming speech-to-speech translation using semantic speech tokens [A/OL]. arXiv: 2410.03298, 2024.
- [99] INAGUMA H, POPURI S, KULIKOV I, et al. UnitY: Two-pass direct speech-to-speech translation with discrete units [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 15655-15680.
- [100] SEAMLESS_COMMUNICATION. Seamless: Multilingual expressive and streaming speech translation [A/OL]. arXiv: 2312.05187, 2023.
- [101] AGRANOVICH A, NACHMANI E, RYBAKOV O, et al. Simultron: On-device simultaneous speech to speech translation [A/OL]. arXiv: 2406.02133, 2024.
- [102] JIA Y, RAMANOVICH M T, REMEZ T, et al.

- Translatotron 2: High-quality direct speech-to-speech translation with voice preservation[C/OL]//Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022; 10120-10134.
- [103] LE C, QIAN Y, WANG D, et al. TransVIP: Speech to speech translation system with voice and isochrony preservation [C]//Proceedings of the 38th Annual Conference on Neural Information Processing Systems, 2024.
- [104] LABIAUSSE T, MAZARÉ L, GRAVE E, et al. High-fidelity simultaneous speech-to-speech translation[A/OL]. arXiv: 2502.03382, 2025.
- [105] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[A]. arXiv: 2302.13971, 2023.
- [106] WANG M, ZHAO J, VU TT, et al. Simultaneous machine translation with large language models C/OL]//Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association. Canberra, Australia, 2024; 89-103.
- [107] WANG M, VU T T, WANG Y, et al. Conversational simulmt: Efficient simultaneous translation with large language models [A/OL]. arXiv: 2402.10552, 2024.
- [108] AGOSTINELLI V, WILD M, RAFFEL M, et al. Simul-LLM: A framework for exploring high-quality simultaneous translation with large language models [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024; 10530-10541.
- [109] RAFFEL M, AGOSTINELLI V, CHEN L. Simultaneous masking, not prompting optimization: A paradigm shift in fine-tuning LLMs for simultaneous translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA, 2024; 18302-18314.
- [110] KOSHKIN R, SUDOH K, NAKAMURA S. TransLLaMA: LLM-based simultaneous translation system [C]//Proceedings of the Association for Computational Linguistics, Miami, Florida, USA, 2024; 461-476.
- [111] YU D, ZHAO Y, ZHU J, et al. SimulPL: Aligning human preferences in simultaneous machine translation [C]//Proceedings of the 38th International Conference on Learning Representations, 2025.
- [112] GUO S, ZHANG S, MA Z, et al. Agent-SiMT: Agent-assisted simultaneous machine translation with large language models[A/OL]. 2024. arXiv: 2406.06910, 2024.
- [113] CHENG S, HUANG Z, KO T, et al. Towards achieving human parity on end-to-end simultaneous speech translation via llm agent [A/OL], 2024. 2407.21646, 2024.
- [114] BOJAR O, CHATTERJEE R, FEDERMANN C, et al. Proceedings of the 10th workshop on statistical machine translation[C]. Proceedings of the Association for Computational Linguistics, 2015.
- [115] DI GANGI M A, CATTONI R, BENTIVOGLI L, et al. MuST-C: A multilingual speech translation corpus [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, 2019; 2012-2017.
- [116] JIA Y, TADMOR RAMANOVICH M, WANG Q, et al. CVSS corpus and massively multilingual speech-to-speech translation [C]//Proceedings of the 13th Language Resources and Evaluation Conference. Marseille, France; European Language Resources Association, 2022; 6691-6703.
- [117] MACHÁČEK D, BOJAR O, DABRE R. MT metrics correlate with human ratings of simultaneous speech translation[C]//Proceedings of the 20th International Conference on Spoken Language Translation. Toronto, Canada, 2023; 169-179.
- [118] KANO Y, SUDOH K, NAKAMURA S. Average token delay: A latency metric for simultaneous translation [C]//Proceedings of the Interspeech, 2023; 4469-4473.
- [119] MAKINAE M, SUDOH K, YAMADA M, et al. An automatic quality metric for evaluating simultaneous interpretation[A/OL]. 2025. arXiv: 2407.06650, 2025.
- [120] NGUYEN H, ESTÈVE Y, BESACIER L. An empirical study of end-to-end simultaneous speech translation decoding strategies [C]//Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 2021; 7528-7532.
- [121] FANG Q, GUO S, ZHOU Y, et al. Llama-omni: Seamless speech interaction with large language models [C]//Proceedings of the 15th International Conference on Learning Representations, 2025.
- [122] GUO S, ZHANG S, MA Z, et al. Large language models are read/write policy-makers for simultaneous generation [A/OL]. arXiv: 2501.00868, 2025.



张绍磊(1998—), 博士研究生, 主要研究领域为自然语言处理、实时翻译和大语言模型。
E-mail: zhangshaolei20z@ict.ac.cn



冯洋(1982—), 通信作者, 博士生导师, 主要研究领域为自然语言处理和大语言模型。
E-mail: fengyang@ict.ac.cn

征稿通知 | 第十一届中国健康信息处理大会(CHIP 2025)

大会介绍

中国健康信息处理大会(China Health Information Processing Conference, CHIP)是中国中文信息学会(CIPS)医疗健康与生物信息处理专业委员会开展的“以信息处理技术助力探索生命之奥秘、提高健康之质量、提升医疗之水平”为主旨的年度会议。CHIP 是中国健康信息处理领域的重要会议,是世界各地学术界、企业界和政府部门的研究人员和从业人士分享创意,进一步推广领域研究成果和经验的重要平台。本次第十一届中国健康信息处理大会(CHIP 2025)将于 2025 年 11 月 14-16 日在广东东莞举行。我们诚挚地邀请大家莅临 CHIP 2025,共襄盛举。

会议网站: <http://www.cips-chip.org.cn/2025/>

论文主题(包括但不限于)

- 医学基础模型、大模型、预训练模型
- 计算健康信息分析
- 健康与医疗的深度学习
- 下一代健康信息处理技术
- 使用新型数据源(例如病例报告、社交媒体或临床研究数据)进行临床决策
- 多模态医疗数据处理
- 跨模态学习
- 多模态学习在医学领域的应用
- 医学知识表示学习与知识工程
- 健康信息处理的机器学习算法
- 病历质控与智能化医疗决策管理
- 人工智能驱动的科学发现

论文投稿

大会诚挚邀请相关领域的学者和研发人员将未经发表的成果投稿英文论文。大会论文集将由 Springer 出版 CCIS 专辑(EI 和 ISTP 索引),优秀论文推荐到多个英文期刊(Health Information Science and Systems(SCI)、Mathematics(SCI)、Intelligent Medical Science 等)和中文期刊《医学信息学》、《数字医学与健康》等。

论文投稿遵循 Springer 标准,论文格式和模版(Word、Latex)可在 Springer 官网下载:

<https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines>

截止日期: 2025 年 09 月 30 日

录用通知: 2025 年 10 月 15 日

终稿提交: 2025 年 10 月 20 日

提交链接: <https://easychair.org/conferences?conf=chip2025>

比赛评测

大会已正式开启评测(详见评测任务),还将颁发正式获奖证书,评审获奖团队还将受邀撰写技术论文,在专委会的指导下在相关期刊上发表,欢迎相关领域的科研人员积极参与评测竞赛。

专题征召

大会将设置若干专题论坛,内容可包括大模型评测、医学具身智能、医疗隐私保护、跨模态学习等,大会提供会务支持,欢迎报名。

(详情见中国中文信息学会官方公众号)

(中国中文信息学会)