Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

机器翻译研究进展

王海峰[a,*], 吴华[a], 何中军[a], 黄亮[b], Kenneth Ward Church[b]

[a] *Baidu Inc., Beijing 100193, China*
[b] *Baidu Research, Sunnyvale, CA 94089, USA*

ARTICLE INFO

Article history:
Received 15 November 2020
Revised 30 January 2021
Accepted 29 March 2021
Available online 14 July 2021

关键词
机器翻译
神经网络机器翻译
同声传译

摘要

经过70多年的发展，机器翻译取得了巨大成就。特别是近年来，随着神经网络机器翻译（NMT）的出现，翻译质量得到了极大提高。本文首先回顾机器翻译的发展历程，从基于规则的机器翻译、基于实例的机器翻译，到统计机器翻译。然后详细介绍神经网络机器翻译技术的进展，包括基本原理和当前主流模型（Transformer），以及多语言翻译。接下来介绍机器同声传译的最新进展，探讨如何在翻译质量和时间延迟方面取得平衡。之后，介绍机器翻译丰富的产品形式和应用。最后，简要讨论机器翻译面临的挑战和未来的研究方向。

1. 机器翻译发展简史

机器翻译（MT）研究如何使用计算机将一种语言翻译成另一种语言。第一台计算机——电子数字积分计算机——问世一年之后，Warren Weaver于1947年首次提出了机器翻译的设想[1]。从那时起，机器翻译就被认为是自然语言处理（NLP）领域中最具挑战性的任务之一。

从方法上来看，机器翻译技术可以分为两大类：基于规则的方法和基于语料库的方法。从机器翻译设想提出到20世纪90年代，基于规则的方法一直占据主导地位。基于规则的机器翻译（RBMT）使用双语词典和人工撰写的规则将源语言文本翻译成目标语言文本。然而，人工撰写规则成本很高，规则维护难度大，很难从一个领域转换到另一个领域，从一种语言转换到另一种语言。因此，基于规则的系统很难扩展到开放领域翻译和多语言翻译。机器翻译发展初期其主要被应用于军事领域。1954年，乔治敦大学与IBM公司合作，首次使用IBM-701计算机完成了将俄语翻译为英语的实验，拉开了机器翻译从梦想走向现实的序幕。之后的十多年里，机器翻译一直是热点研究领域。但随着1966年美国语言自动处理咨询委员会（ALPAC）发表关于机器翻译的报告，这股热潮戛然而止[2]。该报告对机器翻译持怀疑态度，导致机器翻译研究经费大幅削减，相关研究变得极其困难。在机器翻译繁荣发展的1962年，成立了当今计算语言学领域最具影响力的学术组织——国际计算语言学学会（Association for Computational Linguistics, ACL），其成立初期的名字为机器翻译与

* Corresponding author.
E-mail address: wanghaifeng@baidu.com (H. Wang).

英文原文：Engineering 2022, 18(11): 143–153
引用本文：Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, Kenneth Ward Church. Progress in Machine Translation. *Engineering*, https://doi.org/10.1016/j.eng.2021.03.023

计算语言学学会（Association for Machine Translation and Computational Linguistics, AMTCL）。然而到1968年，ALPAC报告发表后，机器翻译发展进入萧条期，该学会将“MT”从其名称中删除。即便是在机器翻译研究遇冷的这段时间，研究人员也一直不断尝试各种方法以提高翻译质量。1965年，自然语言处理领域的研究人员举办了第一届国际计算语言学会议（COLING），会议重点是基于规则的句法分析和翻译。从20世纪70年代开始，RBMT方法变得更加成熟。1978年，SYSTRAN公司推出了商业翻译系统，这是当时基于规则的机器翻译系统取得商业化应用的著名系统之一。谷歌在2007年之前一直使用SYSTRAN公司的机器翻译服务。

随着双语语料库的不断建设和完善，基于语料库的机器翻译逐渐成为主流。其主要有三种方法：基于实例的机器翻译（EBMT）、统计机器翻译（SMT）和神经网络机器翻译（NMT）。20世纪80年代中期，研究人员提出了EBMT方法，其主要思想是通过模仿从双语语料库中检索出的相似例句来实现翻译[3]。EBMT的翻译效果依赖于检索到的例句质量。检索到的例句质量越高、与原文的匹配度越大，翻译效果越好。然而，由于双语语料库难以涵盖所有语言现象，导致EBMT方法在检索相似例句时覆盖率较低，进而影响翻译质量。因此，EBMT方法通常应用于计算机辅助翻译系统，提供相似例句作为翻译参考。

1990年，Brown等[4]提出了SMT方法，其主要思想是机器从大量数据中自动学习翻译知识，而不是依靠人类专家撰写规则。进一步地，在1993年，他们提出了5个SMT模型[5]，形式化地刻画翻译过程。由于SMT方法的复杂性，以及20世纪80~90年代RBMT在商业应用中的主导地位，当时SMT方法并未被广泛采用。然而，统计方法的出现受到学术界的重视。1996年，研究人员发起并召开了第一届自然语言处理中的经验方法会议（EMNLP），其目的是汇集来自一系列不同学科的经验方法，包括语言学中基于语料库的方法和工程学中的信息论[6]。1999年，研究人员在约翰斯·霍普金斯大学举办了一场夏季研讨会[7]。研讨会的成果之一是复现了Brown等提出的5个模型，并发布了一个名为“Egypt”的SMT工具包，大大降低了SMT的研究门槛。随后，词对齐工具GIZA和GIZA++相继发布[8]。2003年，基于短语的SMT方法[9]进一步提高了机器翻译质量。基于此方法的开源系统“Pharaoh”及其升级版本“Moses”[10]极大地促进了SMT系统的发展。基于以上开源工具及系统，SMT方法得到广泛研究和应用。2006年，谷歌推出了以基于短语的SMT为主要系统的互联网翻译服务。微软和百度等公司也在随后几年推出了机器翻译服务。需要注意的是，在实际应用中，单一模型很难解决丰富多样的翻译需求。因此，实际应用中通常采用集成了多种机器翻译模型的混合方法[11]，以提高翻译质量。受SMT模型成功的鼓舞，研究人员提出了多种创新方法来进一步提升SMT的性能，包括引入形态学信息的因子化SMT模型[12]、层次化SMT模型[13]以及在源端和（或）目标端具有句法分析树的基于句法的SMT模型[14–17]。

SMT使用对数线性模型集成多个人工设计的特征，如翻译模型、语言模型和重排序模型等，尽管能够较显著地提升翻译质量，但在处理语序差异大的语言对翻译时仍然面临严重的词语重排序问题。随着深度学习技术在语音处理、计算机视觉等领域的快速发展，研究人员开始将深度学习技术应用于机器翻译。2014年，Bahdanau等[18]和Sutskever等[19]提出了端到端神经网络机器翻译模型，并正式使用了“神经网络机器翻译”（neural machine translation, NMT）一词。NMT的基本思路是将源语言映射成稠密向量（语义表示），然后基于注意力机制生成译文。随后，Dong等[20]提出了一种基于NMT的多语言翻译框架，这被认为是NMT多语言翻译的突破性方法。2015年，百度部署了世界上第一个大规模NMT系统[21]。2016年，谷歌也推出了NMT系统[22]。此后，其他公司陆续发布了NMT系统。自2014年NMT被提出以来，仅用了大约一年的时间就实现了大规模在线部署。相比之下，SMT系统应用于在线服务花了大约16年的时间。此后，基于卷积神经网络的翻译模型[23]和Transformer模型[24]，再次显著提高了NMT系统的翻译质量。NMT的巨大进步甚至引发了关于机器翻译是否可以与人工翻译相媲美的广泛讨论。越来越多的研究围绕NMT展开，如非自回归模型[25–26]、无监督NMT模型[27–28]和NMT预训练模型[29]等，旨在提高多语言翻译质量和翻译效率。

语音处理和机器翻译取得的巨大进步使得语音翻译成为前沿和热点方向。对口语翻译或语音翻译的探索始于1983年国际电信联盟博览会上展示的一个小型实验性自动口译系统[30]。1988年出现的语音到语音（S2S）翻译系统SpeechTrans [31]，被认为是语音翻译中的一个重要里程碑式系统[32]。在随后的20年中，特别是自1991年国际先进语音翻译研究联盟（C-STAR）成立以来，从限定领域和限定词汇的系统[33–35]到开放领域的自然语音翻译[36–40]，语音翻译的发展令人瞩目。2004年，国际口语翻译研讨会（IWSLT）首次举办并延续至今，进一步促进了语音翻译的发展[39]。

随着神经网络技术在机器翻译和语音识别领域的发

展，新的语音翻译系统旨在实现同声传译的自动化，即在低时间延迟（通常只有几秒钟）的情况下，实现与源语言语音（几乎）同步的自动翻译。同声传译对人类来说也是极具挑战性的，需要极高的专注力来倾听和理解源语言，同时需要娴熟的翻译技巧快速地翻译为目标语言并传递给听众。因此，全世界范围内合格的同声传译员数量十分有限。同声传译员通常由两名或更多人组成团队，每15~30 min交替工作，以防止错误率呈指数增长[38]。受短时记忆限制，同声传译员通常采用合理省略源语言内容等翻译技巧[41]，以兼顾翻译准确度与时间延迟。因此，迫切需要开发机器同传技术，以减轻人类同传译员的负担，降低同传成本。作为一项早期工作，Wang等[42]提出了一种基于神经网络的机器同传方法，将流式语音切分成适当的片段以提高语音翻译质量。为了满足机器同传低时延要求，Ma等[43]提出了一种简单有效的“前缀到前缀”的机器同传模型。该技术首次实现了可控时间延迟，重新激发了NLP领域对机器同传的研究兴趣。国际上许多公司，如谷歌、微软、脸书、华为等，纷纷加入这一方向的研究。百度等公司的机器同传系统在数百场会议中得到了实际应用。为了促进相关技术发展，2020年，研究人员在ACL举办了第一届国际机器同传研讨会。同年，IWSLT也开设了新的语音翻译赛道。

2. 神经网络机器翻译

近年来，NMT发展迅速[44–45]。典型的NMT模型包含两部分：编码器将源句子映射为向量，解码器基于该向量生成译文。这个过程类似于人类翻译。NMT模型首先“读取”整个源句子；然后，基于对句子的理解，翻译模型逐词生成目标句子。与RBMT和SMT等以前的方法相比，NMT不需要人工撰写规则和设计特征。NMT是一个端到端的框架，直接从训练语料库中学习语义表示和翻译知识。凭借这些优势，NMT成为机器翻译领域当前的主流方法。

本节首先介绍NMT模型，包括基于基本循环神经网络（RNN）的模型及其改进，以及当前主流的NMT模型Transformer。然后，介绍多语言翻译，并讨论能够充分利用数据的回译技术和基于枢轴语言的翻译技术，以及基于多任务学习的翻译模型与多语言统一翻译模型等。接下来，介绍语音翻译及机器同传最新进展，包括由语音识别（ASR）、机器翻译和语音合成（TTS）组成的级联模型，以及直接对语音和翻译建模的端到端模型。

2.1. 神经网络机器翻译模型

典型的NMT模型是基于标准RNN或其变体构建的[18–19]。给定源句子$x=\{x_1,x_2,\ldots,x_{T_x}\}$（其中，$T_x$表示$x$的长度），编码器将$x$压缩为隐状态$h=\{h_1,h_2,\ldots,h_{T_x}\}$，如下所示：

$$h_t=g(h_{t-1},\ x_t,\theta) \tag{1}$$

式中，$g(.)$是激活函数；h_t和x_t分别是在时间t的隐状态和源语言词向量；t表示时间步长；θ是模型参数。在基本模型中，编码器将最后一个隐状态h_{T_x}作为源句子的表示。然后，解码器根据下式生成译文：

$$p\left(y|x\right)=\prod_{t=1}^{T_y}p(y_t|y_{<t},\boldsymbol{c}) \tag{2}$$

式中，$y=\{y_1,y_2,\ldots,y_{T_y}\}$是目标句子；$p\left(y|x\right)$是翻译概率；$T_y$是$y$的长度；$\boldsymbol{c}$是从隐状态$h$生成的向量；$y_t$是目标词；$y_{<t}=\{y_1,y_2,\ \ldots,y_{t-1}\}$是已经生成的目标词。

标准RNN模型的缺点之一是信息在传递过程中衰减很快，导致长句翻译质量严重下降。为了克服这一问题，Bahdanau等[18]提出了三种改进方案，被广泛应用于NMT模型。接下来逐一介绍。

2.1.1. 注意力机制

当生成目标单词时，与上述基本模型中使用编码器最后一个隐状态h_{T_x}来表示源句子不同，注意力机制计算目标单词和所有源单词之间的关联，并评估关联的强度。

$$\boldsymbol{c}_t=\sum_{j=1}^{T_x}a_{tj}h_j \tag{3}$$

式中，$\boldsymbol{c}_t$是上下文向量；h_j是源单词x_j的隐状态；j是x的单词索引；a_{tj}是目标单词y_t和h_j的关联权重，其计算公式如下：

$$a_{tj}=\frac{\exp\left(e_{tj}\right)}{\sum_{i=1}^{T_x}\exp\left(e_{ti}\right)} \tag{4}$$

式中，e_{tj}是由前馈神经网络计算得到的词对齐强度；i是x的单词索引。

实际上，注意力机制类似于SMT中使用的词对齐。SMT中的词对齐是一种“硬对齐”，表示源单词和目标单词是否有连接。而NMT中的注意力机制是一种“软对齐”，将目标单词通过不同权重连接到所有源单词。注意力机制显著提高了翻译质量，使NMT成为MT历史上的一项突破性技术。

2.1.2. 双向编码

与单向编码从左到右计算隐状态不同，双向编码器根

据从左到右和从右到左两个方向计算隐状态，如$\overrightarrow{h}=\left\{\overrightarrow{h}_1,\overrightarrow{h}_2,\ldots,\overrightarrow{h}_{T_x}\right\}$和$\overleftarrow{h}=\left\{\overleftarrow{h}_1,\overleftarrow{h}_2,\ldots,\overleftarrow{h}_{T_x}\right\}$。然后将隐状态拼接为$h=\left\{\left[\overrightarrow{h}_1,\overleftarrow{h}_1\right],\left[\overrightarrow{h}_2,\overleftarrow{h}_2\right],\ldots,\left[\overrightarrow{h}_{T_x},\overleftarrow{h}_{T_x}\right]\right\}$。因此，对于任意一个时刻，隐状态既包含了此时刻之前的历史信息，也包含了此时刻之后的未来信息，这再次提高了翻译质量。

2.1.3. 门控循环单元

门控循环单元（GRU）是传统简单激活函数的一种变体。GRU类似于长短时记忆网络（LSTM）[46]，但效率更高。GRU和LSTM都允许网络学习长距离依赖关系，而不会受到梯度消失问题的影响[47]。

实验表明，与SMT相比，NMT有显著进步。然而，早期的NMT模型仍然存在缺点，如集外词（OOV）问题、漏译问题、解码速度慢等。为了克服这些问题，He等[48]提出将统计特征（如短语表、n元语言模型和长度惩罚）引入NMT。沿着这个方向，研究人员借鉴了SMT技术，并将其融入NMT中，如词语覆盖度[49]、对齐一致性[50]、句法信息[51–53]、短语表[54–55]和翻译建议[56]等。Sennrich等[57]使用字节对编码（BPE）的压缩算法[58]进行分词，将开放词汇表压缩为固定大小的子词词汇表。该方法简单高效，被广泛用于NMT以解决集外词和低频词翻译问题。

基于RNN的NMT在编解码过程中对当前词的处理依赖于前文信息，难以并行化。针对这一问题，研究人员提出了多种方案以提升NMT模型并行能力。例如，将计算机视觉中常用的卷积神经网络（CNN）引入NMT [23]，通过卷积操作实现对句子中的长距离单词依赖关系高效建模，显著提升了模型的并行化能力。

受基于CNN的NMT方法的启发，Vaswani等[24]提出了一个名为Transformer的新型网络。该网络完全基于注意力机制，没有任何循环和卷积操作。Transformer包含三种注意力：编码器自注意力、解码器掩码注意力和编码器-解码器注意力。研究人员提出了一种新的缩放点积方法来计算这几种注意力。

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{5}$$

式中，$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别是查询向量、键向量和值向量；$\sqrt{d}$是缩放比例因子；$\boldsymbol{K}^{\mathrm{T}}$是$\boldsymbol{K}$的转置。具体来说，对于每个单词，模型通过将词向量与不同的参数矩阵相乘来创建三个向量——查询向量、键向量和值向量。注意力的作用是计算这些值的加权和，传递到下一层。

此外，研究人员还提出了一种多头注意力机制（multi-head attention mechanism）。

$$\text{Multihead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Concat}\left(\text{head}_1,\ldots,\text{head}_M\right)W^O \tag{6}$$

式中，M是头的个数；$\text{head}_m=\text{Attention}\left(\boldsymbol{Q}W_m^{\boldsymbol{Q}},\boldsymbol{K}W_m^{\boldsymbol{K}},\boldsymbol{V}W_m^{\boldsymbol{V}}\right)$ $(1\le m\le M)$表示不同的注意力空间；$W_m^{\boldsymbol{Q}},W_m^{\boldsymbol{K}},W_m^{\boldsymbol{V}},W^O$是参数矩阵。函数$\text{Concat}\left(\text{head}_1,\ldots,\text{head}_M\right)$将所有注意力头拼接在一起。

与循环神经网络和卷积神经网络相比，Transformer具有更强的并行化和表示能力。因此，它不仅在机器翻译任务上取得了最好效果（state-of-the-art），而且在许多其他NLP任务中也有卓越表现。例如，众所周知的双向编码预训练模型BERT [59]和知识增强预训练模型ERNIE [60]，均基于Transformer构建。

上述模型都是自回归模型，在解码时预测当前词需要依赖于已经生成的单词。这限制了模型在解码期间的并行化能力。针对这一问题，Gu等[25]提出了一种非自回归Transformer（NAT），它可以并行化地生成目标序列。

$$p(y|x)=p_L\left(T|x;\phi\right)\cdot\prod_{t=1}^{T}p(y_t|x;\phi) \tag{7}$$

式中，T是目标句子的长度，采用条件分布$p_L\left(T|x;\phi\right)$建立模型；ϕ是模型参数。

与在生成特殊句尾标记（</s>）时停止解码的自回归模型不同，非自回归模型首先使用$p_L\left(T|x;\phi\right)$来预测目标序列的长度。尽管NAT在解码过程中实现了显著的加速，但翻译质量却受到影响。主要原因是NAT没有对单词依赖性进行建模，其对翻译质量的提升非常重要。受解码效率的鼓舞，研究人员提出了许多方法改进非自回归模型，包括知识蒸馏[61]、模仿学习[26]和课程表学习[62]等。

2.2. 多语言翻译

不同的语言具有不同的形态和结构，这使得语言之间的翻译不仅对机器翻译来说是一项艰巨的任务，而且对人类专家而言也同样充满挑战。例如，汉语和英语是主-谓-宾型语言，而日语和韩语是主-宾-谓型语言。在进行汉语和日语之间的翻译时，通常需要进行长距离重新排序。此外，汉语是一种形态变化少的孤立型语言，而日语是一种具有丰富词形变化的黏着型语言。语言之间的差异性增加了多语言机器翻译的难度。

数据驱动的机器翻译方法，无论是SMT还是NMT，从大量平行语料中自动学习翻译知识。一般来说，增加训练数据量能提高翻译质量。Koehn和Knowles [63]的实验表明，当英语-西班牙语翻译的训练词数从40万增加到3.857亿时，翻译质量（使用自动评价指标BLEU度量）提高了约30%（绝对提升）。

遗憾的是，世界上大多数语言缺乏平行语料，这些语言也因此被称为“资源贫乏”型语言。由于数据稀疏性问题，为这些语言构建NMT系统是一个巨大的挑战。根据《互联网世界统计》，全球十大语言（英语、汉语、西班牙语、阿拉伯语、葡萄牙语、印尼语/马来语、法语、日语、俄语和德语）在互联网上的用户数量约占互联网用户总数的77%。其中，英语和汉语用户分别占25.9%和19.4%，而所有其他语言用户的总和仅占23.1%。对于资源丰富型语言，如汉语和英语，可以收集数十亿个句对来训练机器翻译模型；然而，对于资源贫乏型语言对，如汉语-印地语或汉语-斯瓦希里语，只有数千个或更少的句对可用。

此外，部署多语言翻译系统的成本也很高。如果在N种语言之间部署翻译系统，通常需要为每个翻译方向（汉译英和英译汉视为两个翻译方向）都构建翻译模型。N种语言互译则需要构建$N\times(N-1)$个翻译模型。

随着NMT技术发展，研究人员一直在寻求克服上述挑战的方法。一般来说，多语言翻译有两种方法：充分利用数据的方法和改进NMT模型的方法。

针对资源贫乏型语言缺乏训练数据的问题，直观的改进方法是收集尽量多的训练数据，并充分挖掘这些数据的潜力。与平行语料库相比，大量单语语料库更容易获得。在NMT中，单语语料通常可用于数据扩充。一种广泛使用的方法是回译[64–65]，其主要思路是首先在一个小型平行语料库上训练一个标准的NMT模型，然后使用该模型翻译大量单语语料（例如，将目标语言句子翻译为源语言句子），从而生成一个可用于重新训练翻译模型的“伪双语语料库”。在极端情况下，可能根本就没有平行语料库。为了解决该问题，可以使用无监督翻译方法构建仅基于源单语语料库和目标单语语料库的翻译系统。Lample等[66]提出将不同语言的句子映射到相同的隐空间，并通过重构句子来训练翻译模型。Artetxe等[67]使用改进的SMT模型来初始化无监督NMT模型，以进一步提高翻译质量。Song等[29]、Conneau和Lample [68]以及Ren等[69]提出了基于预训练的无监督NMT模型。

多语言翻译的另一个研究方向是充分利用资源丰富型语言来提高资源贫乏型语言的翻译质量。该方法可以追溯到SMT时代。使用最广泛的方法是基于枢轴语言的翻译，即使用资源丰富型语言作为枢轴语言，在资源贫乏型语言对之间建立桥梁[70]。以中德翻译为例，由于有大量的中英和英德平行语料，因此可以选择英语作为枢轴语言。最简单的基于枢轴语言的翻译方法是传递法，它使用两个级联翻译系统[71–72]：源语-枢轴语翻译系统，将源语言句子翻译成枢轴语言句子；以及枢轴语-目标语翻译系统，将枢轴语言句子翻译成目标语言句子。该方法易于实现，在实际系统中得到了广泛应用。缺点是级联系统存在误差传播问题。Wu和Wang [73–74]以及Cohn和Lapata [75]提出了一种三角定位法，通过从源语-枢轴语和枢轴语-目标语翻译模型中引入源语-目标语翻译模型来学习短语级别的翻译知识。

此外，多语言NMT还可以使用统一建模方法，充分利用资源丰富型语言来提高资源贫乏型语言的翻译质量。传统的机器翻译方法需要为每个语言对和每项任务建立单独的翻译模型，而NMT使得在一个统一模型中跨不同任务翻译多种语言成为可能。一般来说，根据源端和目标端语言的数量，可以将该研究分为三类：一对多、多对一和多对多。

Dong等[20]提出了一种用于多语言NMT的多任务学习方法。如图1所示，通过共享编码器共享源语言语义表示，该模型可以在不同语言对之间充分利用源语言语料库。该方法为探索将一种源语言翻译成多个目标语言的问题提供了统一的框架。为了在N个语言之间部署翻译系统，该模型只需要训练一个编码器和N个解码器。Luong等[76]将该框架扩展到多任务，包括翻译、句法分析和图像描述。Zoph和Knight [77]提出了一种多对一的NMT模型，该模型在目标端共享解码器。Firat等[78]使用具有共享注意力机制的不同编码器和解码器进行多对多翻译。

Johnson等[79]提出了一种简单的方法，将所有语言放在一起训练一个统一的编码器-解码器模型，以执行多语言翻译。研究人员在源语言句子开头添加了一个特殊标记，以指示它被翻译成哪种目标语言。该方法允许NMT模型学习多语言共享表示[80]，并且实现简单，无需对NMT模型结构进行修改。考虑到语言的多样性，Tan等[81]将语言分为几个群组，并为每个群组训练单独的NMT模型。

在实际系统中，通常将上述方法结合起来，兼顾翻译效率、部署成本等因素。得益于技术进步，当前的翻译系统可以支持数百个语言之间的翻译。Arivazhagan等[82]提出了一种大规模多语言翻译模型，该模型在超过250亿个句对上训练一个具有超过500亿个参数的单一模型，支持103种语言翻译（以英语作为源语言或者目标语言，与其他102种语言之间的翻译）。Fan等[83]提出了M2M-100模型，使用75亿个句对进行训练，可以支持100种语言互译。

2.3. 同声传译

机器同传的目标是实现兼顾翻译质量和翻译效率的高

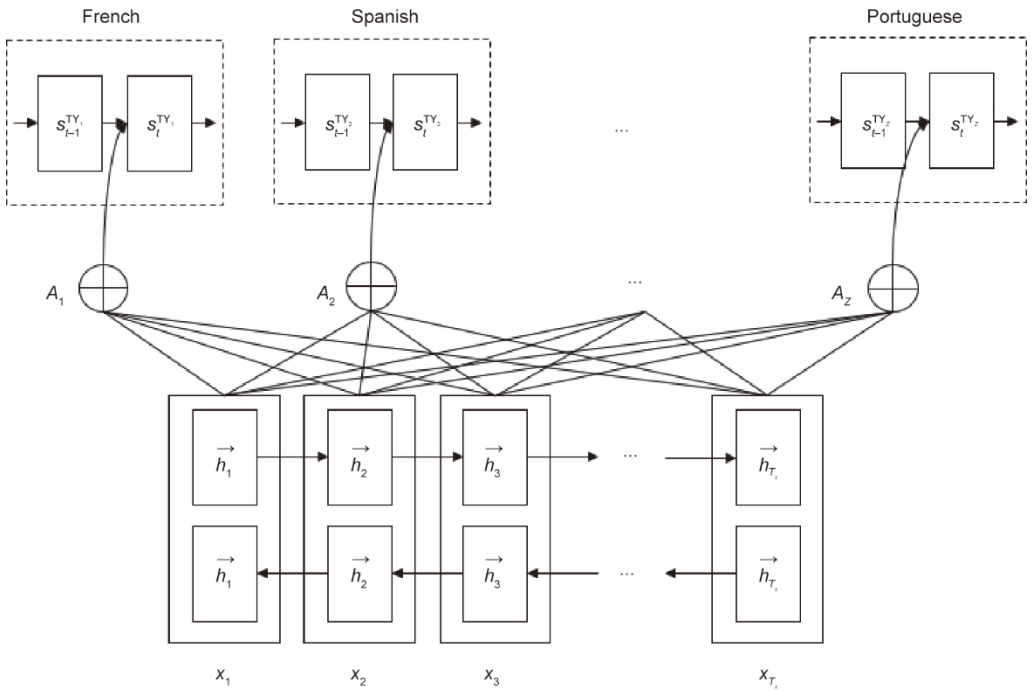

图1. 基于多任务学习的一对多NMT翻译框架图解。$A_1, A_2, \ldots, A_Z$是目标语言的注意力，$\mathrm{TY}_1, \mathrm{TY}_2, \ldots, \mathrm{TY}_Z$是目标语言，$Z$是目标语言数，$s_t^{\mathrm{TY}_z}(1 \leqslant z \leqslant Z)$是解码端的隐状态。

质量实时翻译。在整句翻译（第2.1节）中，机器翻译模型基于整个源语言句子生成目标译文。而在机器同传中，为了保证实时性，翻译模型需要在未得到源语言句子完整内容的条件下进行翻译。

目前，机器同传的研究可以分为两类：级联（流水线）模型和端到端模型。

2.3.1. 级联模型

典型的级联机器同传系统包括将源语音转录为源语言文本流的ASR系统、执行从源文本到目标文本翻译的机器翻译系统，以及生成目标语言语音的TTS系统，具体如图2所示。在实践中，TTS系统是可选的，这取决于不同应用场景中目标端输出的是文本还是语音。

如前所述，机器同传面临的最大挑战是实现高翻译质量和低时间延迟。由于ASR系统输出的文本流没有句子边界，而传统的机器翻译系统将具有明确边界的句子作为输入。因此，ASR的输出与机器翻译的输入不匹配。如果翻译系统在未得到充足的源语言信息之前开始翻译，则翻译质量会降低。反之，如果等待太多的源语言信息，则会增加时间延迟。

为了解决上述问题，需要对ASR的输出进行切分，将切分后的结果作为机器翻译的输入。通常有两种方法：固定文本长度的固定策略和根据上下文动态切分的自适应策略。

固定策略是独立于上下文的预定义的硬策略。此类策略根据固定长度对源文本进行切分[43,84]。Ma等[43]基于“前缀到前缀”的思路提出了wait-k策略，其中，k是模型首先读取的单词数，此后模型边读入边翻译。也就是说，输出总是落后于输入k个单词。该策略受人类同声传译的启发，他们通常在演讲者开始演讲几秒钟后开始翻译，并在演讲结束后的几秒钟内完成翻译。举例而言，如果$k=2$，则使用前两个源词预测第一个目标词，使用前三个源词和生成的第一个目标词预测第二个目标词，依此类推。形式化描述为$y_t: p(y_t|y_{<t}, x_{\leqslant q(t)})$，即使用源语言句子前缀$\{x_1, x_2, \ldots, x_{q(t)}\}$而不是整个源句子来预测目标词。其中，$q(t)$是一个单调非递减函数，表示预测$y_t$时编码器处理的源词数。一般情况下，$q(t)$可以用来表示任意长度的同传策略，其中对于所有$t$，$0 \leqslant q(t) \leqslant |x|$。两种特殊情况除外：①$q(t)=|x|$，此时翻译模型即是传统的整句翻译模型；②$q(t)=0$，则翻译模型退化为一个预测模型，即不依赖源语言句子的任何信息就开始翻译。固定策略简单易

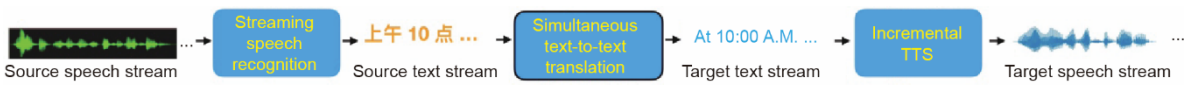

图2. 级联机器同传系统框架。

行，但由于缺乏上下文信息，通常会导致翻译质量下降。

自适应策略根据上下文信息进行动态的源文本切分。通常有两种方式，使用独立的模型对源语言文本流进行切分[85–89]，或者在端到端框架中联合学习切分和翻译[90–91]。自适应策略比固定策略更灵活，取得了更好的效果。受到人类同声传译员翻译方式的启发，Zhang 等[92]提出了一种语义单元驱动的机器同传方法，将源语言文本流动态切分为可独立翻译的片段，以同时满足高质量和低时延要求。

在语音翻译中，有关增量TTS的研究不多。当前主流的TTS系统获取完文本中的所有单词后才开始生成语音，导致时间延迟高。在机器同传中，为了减少延迟，需要以增量方式生成语音。传统的增量TTS方法基于隐马尔可夫模型[93–97]，使用语言特征的完整上下文，每个特征需要单独训练和调参。最近的研究利用了神经网络的优势[98–99]。Yanagita 等[98]提出了一种基于分段的TTS，一次合成一个分段。Ma 等[99]提出了一种神经增量词级TTS。如图3所示，该方法基于两个前提：①单词依赖关系是非常局部的；②音频播放本质上是顺序的，可以与音频生成同时进行。也就是说，可以在合成后续文本时播放已经生成的上一段音频。综上所述，该方法在收到前两个单词后开始生成第一个单词的频谱图；该频谱图被送到声码器以生成第一个单词的波形，该波形会被立即播放。

级联模型易于实现，但是也存在问题。例如，级联系统中的三个模块均需满足实时性要求。此外，ASR错误会在向下游任务传播的过程中被放大，一个单词识别错误可能会导致整体的翻译结果不可接受。因此，需要增强语音翻译系统的健壮性。

2.3.2. 端到端模型

机器同传的最终目标是开发端到端的语音翻译系统，以便源语言语音可以直接翻译成目标语言，而无需像级联方法那样经过中间阶段。端到端模型不仅可以减少级联模型中的错误传播，还可以提高效率。然而，构建高实时性的端到端语音翻译模型是极具挑战性的。此外，可用于训练端到端模型的语音翻译数据非常稀缺。目前，公开可用的机器同传训练数据仅包含数百小时的演讲，其中大部分是日语-英语[100–101]以及欧洲语言[102–103]之间的数据。对于中英翻译，百度发布了一个包含70 h演讲的开放数据集，包括相应的语音转录和翻译[104]。

将语音识别和机器翻译集成到一个统一的框架中并非易事，端到端语音翻译是一项前沿技术。Bansal 等[105]首次验证了端到端语音翻译可以在不用源语言语音转录的情况下实现。近来有些研究基于预训练或多任务学习来提高语音翻译质量。例如，基于ASR数据预训练编码器[105]，利用文本翻译来改进语音翻译[106–108]等。Liu 等[109]使用知识蒸馏方法，通过从机器翻译模型迁移知识来改进端到端语音翻译。但是，这些方法中的不同任务之间不能相互共享信息。为了解决这个问题，研究人员提出了两阶段模型[110–112]，其中第一阶段执行语音识别任务，其隐状态（而非识别结果）作为第二阶段解码器（翻译系统）的输入。Liu 等[113]提出一种交互式端到端语音翻译模型，可以交互地进行语音识别和机器翻译，从而提高了这两项任务的性能。最近也有一些研究聚焦直接建立端到端语音翻译模型[114–115]。然而，由于训练数据有限，以及将语音识别和机器翻译集成到统一框架中的复杂性，目前的端到端语音翻译系统的性能尚不能满足实际要求。

由于级联模型易于部署且翻译质量比较高，因此当前

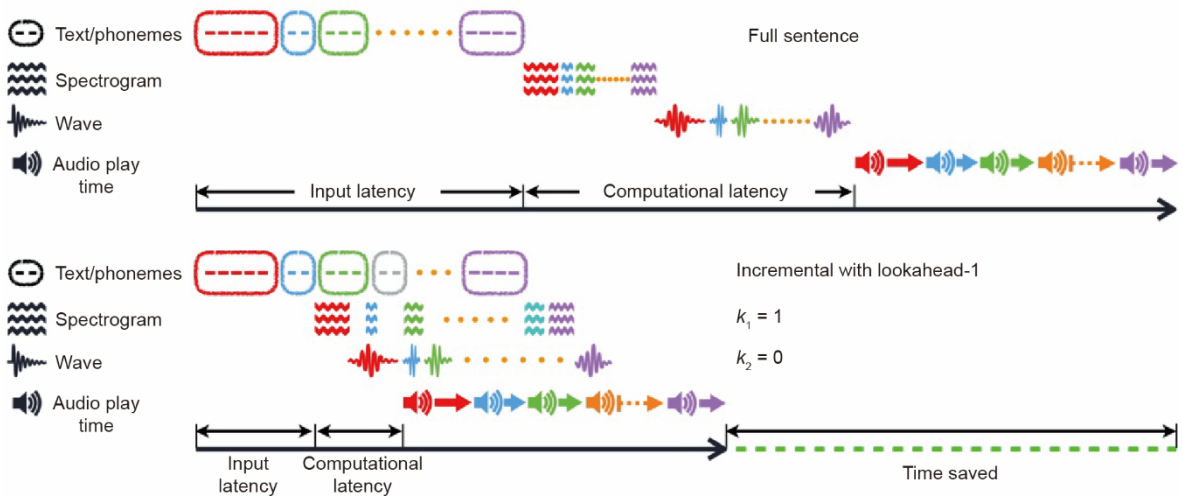

图3. 整句TTS与增量TTS。K_1和K_2分别是频谱图和声波生成的前瞻窗口大小。

大多数实用的语音翻译系统使用该方法。Xiong等[104]将机器同传系统与具有3~7年经验的人类同传译员进行了比较。实验发现，同传译员通常会忽略不重要的信息以保持合理的时间延迟。这可能会损失译文的完整度，但保证了实时性。与同传译员相比，机器同传系统生成的译文完整度更好。Shimizu等[100]实验也表明经验较少的同传口译员在同传过程中会丢失细节。这些研究表明，同声传译对于人类和机器来说都是一项艰巨的任务。

3. 机器翻译应用

机器翻译因其低成本、高效率和高翻译质量而在许多领域得到广泛应用。在中国，人工翻译费用通常为0.1~0.5元/字不等，具体取决于翻译人员的经验丰富程度。而机器翻译的价格约为0.00005元/字符。百度翻译目前支持200多种语言互译，每天翻译量超过千亿字符，应用领域广泛。图4列出了8个较大的领域分布。

3.1. 文本翻译

文本翻译是最常见的机器翻译应用形式。以下是文本翻译的一些典型应用。

（1）网页翻译。随着全球化的迅速发展，快速获取外语信息的需求日益增加。聘请人工翻译人员翻译大量网页既昂贵又耗时。机器翻译提供了一种查看外语网页的便捷方式。用户只需复制/粘贴网页内容或输入网址即可以用母语阅读页面。

（2）科技文献翻译。研究人员、工程师和研究生等用户经常使用机器翻译系统阅读论文和专利等科技文献，或将他们的工作成果翻译成其他语言。例如，为了抗击新型冠状病毒肺炎（COVID-19），生物医学领域的翻译需求迅速增长。科技文献通常包含许多术语。借助领域自适应技术，翻译模型首先使用大规模语料进行预训练，然后使用少量领域内数据进行微调以进一步提升翻译质量。此外，文档翻译用于翻译格式丰富的文档，例如，PowerPoint、Excel、Word和PDF，在生成译文的同时保留字体大小和字体颜色等格式信息。

（3）电子商务翻译。机器翻译广泛用于国际贸易。在机器翻译系统的帮助下，卖家可以快速将网站、产品信息和服务手册翻译成外语，而买家可以轻松购买来自世界各地的产品。此外，机器翻译还可以用于客户服务，以提高服务质量和效率。

（4）语言学习。目前的机器翻译系统通常提供丰富的功能，包括翻译、高质量词典、例句等。因此，用户可以方便地查询单词或短语的含义并学习如何使用它。学生用户经常输入整个段落以帮助阅读理解，并使用例句来辅助写作。

除了文本翻译，基于人工智能技术的最新进展，图像翻译和语音翻译也已广泛应用于实际场景中。

3.2. 图像翻译

图像翻译结合了计算机视觉和机器翻译技术，将图像作为输入，然后将其翻译成目标语言。

（1）多语言图像描述。此类系统可以描述图片内容并进行视觉问答，近年来得到了广泛研究[116–118]。多语言图像描述基于NMT思想，其中，编码器的输入是图像，解码器的输出是文本。由于模型可以为同一张图片生成不同的语言，因此此功能对语言学习非常有帮助。

（2）光学字符识别（OCR）翻译。此种形式的机器翻译首先识别图片中的字符，然后进行翻译并使用译文替换原文本。此功能可用于出国旅行时翻译菜单、街道路牌、产品描述等。随着近年来对文档图像布局和文本信息进行联合建模的研究不断进步[119]，OCR翻译还可用于

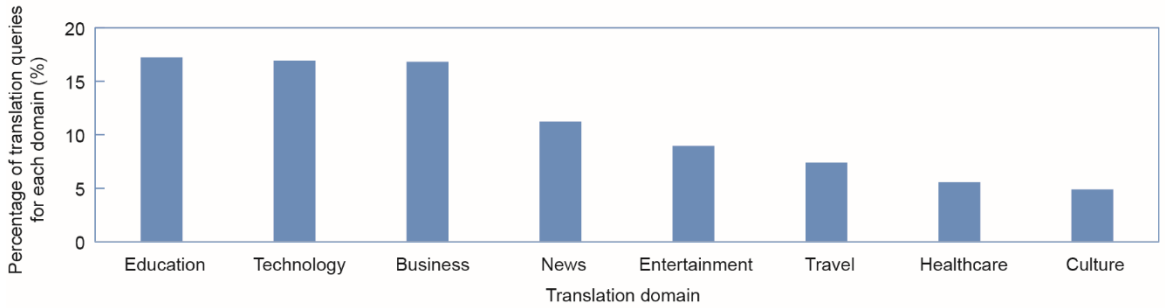

图4. 百度翻译领域分布。

翻译扫描的文档，同时保留原始格式信息。

3.3. 语音翻译

语音翻译结合了语音处理和机器翻译技术，将源语言语音作为输入，并以目标语言文本或语音作为输出。

（1）机器同声传译。如第2.3节所述，机器同传最近取得较大进展，并得到广泛应用。语音到文本（S2T）翻译将语音识别结果和译文以字幕形式投影到屏幕上，以方便用户观看。但是，屏幕上有限的空间通常只能显示一种语言对的字幕。因此，很难将S2T扩展到多语言。语音到语音翻译使得观众可以通过手机收听目标语言声音来解决这个问题。来自不同国家的用户可以选择他们的母语或他们喜欢的任何其他语言。机器同传系统目前广泛应用于国际会议。受新冠疫情影响，越来越多的会议以在线会议的形式举办。针对这类需求，机器同传系统也已集成到在线会议系统中，提供实时翻译。此外，用户可以使用机器同传插件用母语观看外语视频，如电影和讲座等。

（2）便携式翻译设备。带有语音翻译功能的移动设备近年来受到用户青睐。它们易于携带和使用，在语言学习、海外旅行和商务谈判等许多场景中有广泛应用。

此外，机器翻译技术也可用于诗歌生成[120]和中文对联生成。以诗歌生成为例，机器翻译模型将前一行生成的诗句作为“源语言句子”，将后续诗句作为“目标语言句子”，则可以逐行生成诗歌。

4. 挑战和展望

尽管当前机器翻译取得了显著进步，但仍有很大的提升空间。在机器翻译研讨会（WMT）等开展的机器翻译评测中，某些基准测试集上的自动评价指标（如BLEU、WER、METEOR等）[121–123]表明，机器翻译有时比人工翻译更好。但需要注意的是，这些指标很难全面反映译文质量。好的翻译至少应该具备两个基本特点：译文忠实于原文（忠实度），以及译文地道流畅（流利度）。NMT方法在某些语言对或者领域翻译中表现出较高的忠实度和流利度。然而，该方法远非完美，在有些任务如语音翻译上，仍面临较大挑战。

总体而言，机器翻译还有许多方面有待改进。

第一，需要设计新的评价指标来衡量机器译文不同部分的重要程度。例如，人类同传译员在进行同声传译时不会试图翻译所有内容。在同传过程中，知道哪些内容需要翻译以及何时开始翻译是非常重要的。同传译员知道何时需要加快速度，何时可以放缓节奏；知道哪些内容需要着重强调，哪些内容则可以省略不译。但是，机器同传系统会翻译所有内容，并且不知道如何省略非重要内容以减少时间延迟。进一步地，机器同传系统应该反映出演讲者所强调的重点内容。最近，有些研究使用声学特征来识别重点内容并将其翻译成目标语言[124–126]。除了语音信息外，说话者的肢体语言和韵律也可以清晰传达说话者所强调的某一部分内容（相对于其他部分而言）。然而，将翻译与说话者的肢体语言同步是比较困难的。此外，演讲者在演讲时经常会参考幻灯片。同样地，将翻译与幻灯片内容同步也充满挑战。尽管BLEU和WER之类的评价指标能够一定程度上衡量译文的完整性，但是不够全面，没有涉及延迟、强调、同步、理解等，这些也是影响翻译的重要因素。在机器同传中，前端ASR系统不仅需要能识别单词，还应该能够识别说话人所强调的重点内容，这些内容将会影响下游任务（机器翻译、语音合成）的效果。因此，新的评价指标应该奖励同传系统将重要内容做出准确翻译，同时惩罚只将非重点内容做出翻译。

第二，机器翻译的鲁棒性需要进一步提高。有时源句子的微小改变（如词语或标点符号的改变）可能会导致机器翻译产生的译文发生巨大变化。与机器相比，人类具有很强的容错能力，能够灵活地处理各种非标准语言现象和错误，有时甚至下意识地予以纠正。高鲁棒性的机器翻译系统在实际应用中至关重要。研发可解释的机器翻译系统是一种可能的解决方案。

第三，NMT在资源贫乏的语言对和领域中面临着严重的数据稀疏问题。目前的机器翻译系统通常使用数千万甚至数亿个句对的数据进行训练，从而获得较高的翻译质量。数据稀缺会导致机器翻译质量变差。与机器相比，人类却能从少量样本中学习。尽管研究者已经提出了多种数据增强方法、多任务学习方法和预训练方法来缓解多语言翻译面临的数据稀疏问题，但如何提高资源贫乏型语言的翻译质量仍任重道远。

综上所述，要实现高质量的机器翻译还有很长的路要走。需要研发能够结合符号规则、知识和神经网络的新方法，以进一步提高翻译质量。幸运的是，机器翻译在实际场景中的广泛应用可以不断提供更多更丰富的数据，促进机器翻译新方法的快速发展。

Compliance with ethics guidelines

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church declare that they have no con-

flict of interest or financial conflicts to disclose.

References

[1] Weaver W. Translation. Mach Transl Lang 1955;14:15–23.

[2] Hutchins J. ALPAC: the (in) famous report. In: Nirenburg S, Somers HL, Wilks YA, editors. Readings in machine translation. Cambridge: MIT Press; 2003.

[3] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle. In: Elithorn A, Banerji R, editors. In: Proceedings of the International NATO Symposium on Artificial and Human Intelligence. New York City: Elsevier North-Holland, Inc; 1984. p. 173–80.

[4] Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty JD, et al. A statistical approach to machine translation. Comput Linguist 1990;16(2): 79–85.

[5] Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL. The mathematics of statistical machine translation: parameter estimation. Comput Linguist 1993; 19(2):263–311.

[6] Church KW, Mercer RL. Introduction to the special issue on computational linguistics using large corpora. Comput Linguist 1993;19(1):1–24.

[7] Al-Onaizan Y, Curin J, Jahr M, Knight K, Lafferty J, Melamed D, et al. Statistical machine translation: final report. Baltimore: Johns Hopkins University Summer Workshop; 1999.

[8] Och FJ, Ney H. A systematic comparison of various statistical alignment models. Comput Linguist 2003;29(1):19–51.

[9] Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; 2003 May 27–Jun 1; Edmonton, AB, Canada; 2003.

[10] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, et al. Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.

[11] Wang H. [Multi-strategy machine translation]. In: Cao YQ, Sun MS, editors. [Frontiers of Chinese information processing]. Beijing: Tsinghua University Press; 2006. p. 45–52. Chinese.

[12] Koehn P, Hoang H. Factored translation models. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CoNLL); 2007 Jun 25–27; Prague, Czech Republic; 2007.

[13] Chiang D. Hierarchical phrase-based translation. Comput Linguist 2007;33(2): 201–28.

[14] Yamada K, Knight K. A syntax-based statistical translation model. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics; 2001 Jul 6–11; Toulouse, France; 2001.

[15] Galley M, Graehl J, Knight K, Marcu D, DeNeefe S, Wang W, et al. Scalable inference and training of context-rich syntactic translation models. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics; 2006 Jul 17–21; Sydney, NSW, Australia; 2006.

[16] Liu Y, Liu Q, Lin S. Tree-to-string alignment template for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics; 2006 Jul 17–21; Sydney, NSW, Australia; 2006.

[17] Graehl J, Knight K, May J. Training tree transducers. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; 2004 May 2–7; Boston, MA, USA; 2004.

[18] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, USA; 2015.

[19] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems; 2014 Dec 8–13; Montreal, QC, Canada; 2014.

[20] Dong D, Wu H, He W, Yu D, Wang H. Multi-task learning for multiple language translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015 Jul 26–31; Beijing, China; 2015.

[21] Pouliquen B. WIPO Translate: patent neural machine translation publicly available in 10 languages [presentation]. In: Machine Translation XVI; 2017 Sep 18–22; Nagoya, Japan; 2017.

[22] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. arXiv: 1609.08144.

[23] Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia; 2017.

[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA; 2017.

[25] Gu J, Bradbury J, Xiong C, Li VOK, Socher R. Non-autoregressive neural machine translation. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.

[26] Wei B, Wang M, Zhou H, Lin J, Xie J, Sun X. Imitation learning for nonautoregressive neural machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.

[27] Lample G, Conneau A, Denoyer L, Ranzato M. Unsupervised machine translation using monolingual corpora only. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.

[28] Artetxe M, Labaka G, Agirre E. An effective approach to unsupervised machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.

[29] Song K, Tan X, Qin T, Lu J, Liu TY. Mass: masked sequence to sequence pretraining for language generation. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA; 2019.

[30] Kato Y. The future of voice-processing technology in the world of computers and communications. Pro Natl Acad Sci USA 1995;92(22):10060–3.

[31] Tomita M, Tomabechi H, Saito H. SpeechTrans: an experimental real-time speech-to-speech translation. Lang Res 1990;26(4):663–72.

[32] Kitano H. Speech-to-speech translation: a massively parallel memory-based approach. Boston: Kluwer Academic Publishers; 1994.

[33] Waibel A, Jain AN, McNair AE, Saito H, Hauptmann AG, Tebelskis J. JANUS: a speech-to-speech translation using connectionist and symbolic processing strategies. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing; 1991 Apr 14–17; Toronto, ON, Canada; 1991.

[34] Morimoto T, Takezawa T, Yato F, Sagayama S, Tashiro T, Nagata M, et al. ATR's speech translation system: ASURA. In: Proceedings of the 3rd European Conference on Speech Communication and Technology; 1993 Sep 22–25; Berlin, Germany; 1993.

[35] Roe DB, Pereira FCN, Sproat RW, Riley MD, Moreno PJ, Macarron A. Efficient grammar processing for a spoken language translation system. In: Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing; 1992 Mar 23–26; San Francisco, CA, USA; 1992.

[36] Sumita E, Shimizu T, Nakamura S. NICT-ATR speech-to-speech translation system. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.

[37] Fügen C, Kolss M, Paulik M, Stüker S, Schultz T, Waibel A. Open domain speech translation: from seminars and speeches to lectures. In: Proceedings of TC-STAR Workshop on Speech-to-Speech Translation; 2006 Jun 19–21; Barcelona, Spain; 2006.

[38] Moser-Mercer B, Künzli A, Korac M. Prolonged turns in interpreting: effects on quality, physiological and psychological stress (pilot study). Interpreting 1998;3(1):47–64.

[39] Wang H, Wu H, Hu X, Liu Z, Li J, Ren D, et al. The TCH machine translation system for IWSLT 2008. In: Proceedings of International Workshop on Spoken Language Translation; 2008 Oct 20–21; Honolulu, HI, USA; 2008.

[40] Nakamura S, Markov K, Nakaiwa H, Kikui G, Kawai H, Jitsuhiro T, et al. The ATR multilingual speech-to-speech translation system. IEEE Trans Audio Speech Lang Process 2006;14(2):365–76.

[41] He H, Boyd-Graber J, Daume H III. Interpretese vs. translationese: the uniqueness of human strategies in simultaneous interpretation. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA; 2016.

[42] Wang H, Gao W, Li S. Utterance segmentation of spoken Chinese. Chin J Comput 1999;22(10):1009–13. Chinese.

[43] Ma M, Huang L, Xiong H, Zheng R, Liu K, Zhang B, et al. STACL: simultaneous translation with implicit anticipation and controllable latency

using prefixto- prefix framework. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy;2019.
[44] Zhang JJ, Zong CQ. Neural machine translation: challenges, progress and future. Sci China Technol Sci 2020;63(10):2028–50.
[45] Edunov S, Ott M, Auli M, Grangier D. Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31–Nov 4; Brussels, Belgium; 2018.
[46] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9(8):1735–80.
[47] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 1994;5(2):157–66.
[48] He W, He Z, Wu H, Wang H. Improved neural machine translation with SMT features. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence; 2016 Feb 12–17; Phoenix, AZ, USA; 2016.
[49] Tu Z, Lu Z, Liu Y, Liu X, Li H. Modeling coverage for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany; 2016.
[50] Cheng Y, Shen S, He Z, He W, Wu H, Sun M, et al. Agreement-based joint training for bidirectional attention-based neural machine translation. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence; 2016 Jul 9–15; New York City, NY, USA; 2016.
[51] Sennrich R, Haddow B. Linguistic input features improve neural machine translation. In: Proceedings of the First Conference on Machine Translation; 2016 Aug 11–12; Berlin, Germany; 2016.
[52] Wu S, Zhou M, Zhang D. Improved neural machine translation with source syntax. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017 Aug 19–25; Melbourne, VIC, Australia; 2017.
[53] Li J, Xiong D, Tu Z, Zhu M, Zhang M, Zhou G. Modeling source syntax for neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017 Jul 30–Aug 4; Vancouver, CB, Canada; 2017.
[54] Feng Y, Zhang S, Zhang A, Wang D, Abel A. Memory-augmented neural machine translation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017 Sep 7–11; Copenhagen, Denmark; 2017.
[55] Zhao Y, Wang Y, Zhang J, Zong C. Phrase table as recommendation memory for neural machine translation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden; 2018.
[56] Wang X, Lu Z, Tu Z, Li H, Xiong D, Zhang M. Neural machine translation advised by statistical machine translation. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017 Feb 4–9; San Francisco, CA, USA; 2017.
[57] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany; 2016.
[58] Gage P. A new algorithm for data compression. C Users J 1994;12(2):23–38.
[59] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minnesota, MN, USA; 2019.
[60] Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, et al. ERNIE 2.0: a continual pretraining framework for language understanding. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA; 2020.
[61] Zhou C, Neubig G, Gu J. Understanding knowledge distillation in nonautoregressive machine translation. 2019. arXiv:1911.02727.
[62] Guo J, Tan X, Xu L, Qin T, Chen E, Liu TY. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA; 2020.
[63] Koehn P, Knowles R. Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation; 2017 Aug 4; Vancouver, CB, Canada; 2017.
[64] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7−12; Berlin, Germany; 2016.
[65] Poncelas A, Shterionov D, Way A, Wenniger GMDB, Passban P. Investigating backtranslation in neural machine translation. 2018. arXiv:1804.06189.
[66] Lample G, Conneau A, Denoyer L, Ranzato M. Unsupervised machine translation using monolingual corpora only. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3;Vancouver, BC, Canada; 2018.
[67] Artetxe M, Labaka G, Agirre E. An effective approach to unsupervised machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.
[68] Conneau A, Lample G. Cross-lingual language model pretraining. In: Proceedings of the 33rd Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada; 2019.
[69] Ren S, Wu Y, Liu S, Zhou M, Ma S. Explicit cross-lingual pre-training for unsupervised machine translation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China; 2019.
[70] Wang H, Wu H, Liu Z. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In: Proceedings of the COLING/ACL2006 Main Conference Poster Sessions; 2006 Jul 17–21; Sydney, NSW, Australia; 2006.
[71] Utiyama M, Isahara H. A comparison of pivot methods for phrase-based statistical machine translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics; 2007 Apr 22–27; Rochester, NY, USA; 2007.
[72] Khalilov M, Costa-Jussà MR, Henríquez CA, Fonollosa JAR, Hernández A, Mariño JB, et al. The TALP & I2R SMT systems for IWSLT 2008. In: Proceedings of the International Workshop on Spoken Language Translation; 2008 Oct 20–21; Honolulu, HI, USA; 2008.
[73] Wu H, Wang H. Pivot language approach for phrase-based statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.
[74] Wu H, Wang H. Revisiting pivot language approach for machine translation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language; 2009 Aug 2–7; Singapore; 2009.
[75] Cohn T, Lapata M. Machine translation by triangulation: making effective use of multi-parallel corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.
[76] Luong MT, Le QV, Sutskever I, Vinyals O, Kaiser L. Multi-task sequence to sequence learning. In: Proceedings of the International Conference on Learning Representations; 2016 May 2–4; San Juan, Puerto Rico; 2016.
[77] Zoph B, Knight K. Multi-source neural translation. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA; 2016.
[78] Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA; 2016.
[79] Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. Trans Assoc Comput Linguist 2017;5:339–51.
[80] Kudugunta S, Bapna A, Caswell I, Firat O. Investigating multilingual NMT representations at scale. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China; 2019.
[81] Tan X, Chen J, He D, Xia Y, Qin T, Liu TY. Multilingual neural machine translation with language clustering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China; 2019.
[82] Arivazhagan N, Bapna A, Firat O, Lepikhin D, Johnson M, Krikun M, et al. Massively multilingual neural machine translation in the wild: findings and challenges. 2019. arXiv:1907.05019.
[83] Fan A, Bhosale S, Schwenk H, Ma Z, El-Kishky A, Goyal S, et al. Beyond Englishcentric multilingual machine translation. 2020. arXiv:2010.11125.
[84] Dalvi F, Durrani N, Sajjad H, Vogel S. Incremental decoding and training methods for simultaneous translation in neural machine translation. 2018. arXiv: 1806.03661.
[85] Sridhar VKR, Chen J, Bangalore S, Ljolje A, Chengalvarayan R. Segmentation

strategies for streaming speech translation. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2013 Jun 9–14; Atlanta,GA; USA; 2013.

[86] Oda Y, Neubig G, Sakti S, Toda T, Nakamura S. Optimizing segmentation strategies for simultaneous speech translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014 Jun 23–25; Baltimore, MD, USA; 2014.

[87] Cho K, Esipova M. Can neural machine translation do simultaneous translation? 2016. arXiv:1606.02012.

[88] Gu J, Neubig G, Cho K, Li VOK. Learning to translate in real-time with neural machine translation. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; 2017 Apr 3–7; Valencia, Spain; 2017.

[89] Fujita T, Neubig G, Sakti S, Toda T, SimpleNakamura S., lexicalized choice of translation timing for simultaneous speech translation. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association; 2013 Aug 25–29; Lyon, France; 2013.

[90] Arivazhagan N, Cherry C, Macherey W, Chiu CC, Yavuz S, Pang R, et al. Monotonic infinite lookback attention for simultaneous machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.

[91] Ma X, Pino J, Cross J, Puzon L, Gu J. Monotonic multihead attention. In: Proceedings of the International Conference on Learning Representations; 2020 Apr 26–May 1; Addis Ababa, Ethiopia; 2020.

[92] Zhang R, Zhang C, He Z, Wu H, Wang H. Learning adaptive segmentation policy for simultaneous translation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16–20; online; 2020.

[93] Baumann T. Partial representations improve the prosody of incremental speech synthesis. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association; 2014 Sep 14–18;Singapore; 2014.

[94] Baumann T. Decision tree usage for incremental parametric speech synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; 2014 May 4–9; Florence, Italy; 2014.

[95] Pouget M, Hueber T, Bailly G, Baumann T. HMM training strategy for incremental speech synthesis. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association; 2015 Sep 6–10; Dresden, Germany; 2015.

[96] Pouget M, Nahorna O, Hueber T, Bailly G. Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis. In: Proceedings of Inter-Speech 2016; 2016 Sep 8–12; San Francisco, CA, USA; 2016.

[97] Yanagita T, Sakti S, Nakamura S. Incremental TTS for Japanese language. In: Proceedings of Interspeech; 2018 Sep 2–6; Hyderabad, India; 2018.

[98] Yanagita T, Sakti S, Nakamura S. Neural iTTS: toward synthesizing speech in real-time with end-to-end neural text-to-speech framework. In: Proceedings of the 10th ISCA Speech Synthesis Workshop; 2019 Sep 20–22; Vienna, Austria; 2019.

[99] Ma M, Zheng B, Liu K, Zheng R, Liu H, Peng K, et al. Incremental textto-speech synthesis with prefix-to-prefix framework. In: Findings of the Association for Computational Linguistics: EMNLP 2020; 2020 Nov 16–20; online; 2020.

[100] Shimizu H, Neubig G, Sakti S, Toda T, Nakamura S. Collection of a simultaneous translation corpus for comparative analysis. In: Proceedings of Ninth International Conference on Language Resources and Evaluation; 2014 May 26–31; Reykjavik, Iceland; 2014.

[101] Toyama H, Ryu K, Matsubara S, Kawaguchi, Nobuo K, Inagaki Y. CIAIR simultaneous interpretation corpus. In: Proceedings of Oriental COCOSDA; 2004 Nov 17–19; New Delhi, India; 2004.

[102] Sandrelli A, Bendazzoli C. Tagging a corpus of interpreted speeches: the European parliament interpreting corpus (EPIC). In: Proceedings of LREC; 2006 May 22–28; Genoa, Italy; 2004.

[103] Di Gangi MA, Cattoni R, Bentivogli L, Negri M, Turchi M. MuST-C: a multilingual speech translation corpus. In: Proceedings of 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minnesota, MN, USA; 2019.

[104] Xiong H, Zhang R, Zhang C, He Z, Wu H, Wang H. DuTongChuan: context–aware translation model for simultaneous interpreting. 2019. arXiv: 1907. 12984.

[105] Bansal S, Kamper H, Livescu K, Lopez A, Goldwater S. Pre-training on highresource speech recognition improves low-resource speech-to-text translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics; 2018 Jun 1–6; New Orleans, LA, USA; 2018.

[106] Weiss RJ, Chorowski J, Jaitly N, Wu Y, Chen Z. Sequence-to-sequence models can directly translate foreign speech. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association; 2017 Aug 20–24; Stockholm, Sweden; 2017.

[107] Anastasopoulos A, Chiang D. Leveraging translations for speech transcription in low-resource settings. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association; 2018 Sep 2–6; Hyderabad, India; 2018.

[108] Bérard A, Pietquin O, Servan C, Besacier L, Servan C. Listen and translate: a proof of concept for end-to-end speech-to-text translation. In: Proceedings of the 30th Conference on Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain; 2016.

[109] Liu Y, Xiong H, Zhang J, He Z, Wu H, Wang H, et al. End-to-end speech translation with knowledge distillation. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association; 2019 Sep 15–19; Graz, Austria; 2019.

[110] Kano T, Sakti S, Nakamura S. Structured based curriculum learning for end to-end English–Japanese speech translation. In: Proceedings of the 18th Annul Conference of the International Speech Communication Association;2017 Aug 20–24; Stockholm, Sweden; 2017.

[111] Anastasopoulos A, Chiang D. Tied multitask learning for neural speech translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018 Jun 1–6; New Orleans, Louisiana; 2018.

[112] Sperber M, Neubig G, Niehues J, Waibel A. Attention-passing models for robust and data-efficient end-to-end speech translation. Transl Assoc Comput Linguist 2019;7:313–25.

[113] Liu Y, Zhang J, Xiong H, Zhou L, He Z, Wu H, et al. Synchronous speech recognition and speech-to-text translation with interactive decoding. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA; 2020.

[114] Jia Y, Weiss RJ, Biadsy F, Macherey W, Johnson M, Chen Z, et al. Direct speech-to-speech translation with a sequence-to-sequence model. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019); 2019 Sep 15–19; Graz, Austria; 2019.

[115] Kano T, Sakti S, Nakamura S. Transformer-based direct speech-to-speech translation with transcoder. In: Proceedings of the IEEE Spoken Language Technology Workshop; 2021 Jan 19–22; Shenzhen, China; 2021.

[116] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 8–10; Boston, MA, USA; 2015.

[117] Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA; 2017.

[118] Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom–up and top–down attention for image captioning and visual question answering. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018.

[119] Xu Y, Li M, Cui L, Huang S, Wei F, Zhou M. LayoutLM: pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Minning; 2020 Aug 23–27; online; 2020.

[120] Wang Z, He W, Wu H, Wu H, Li W, Wang H, et al. Chinese poetry generation with planning based neural network. Proceedings of the 26th International Conference on Computational Linguistics; 2016 Dec 11–16; Osaka, Japan;2016.

[121] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul 7–12;Philadelphia, PA, USA; 2002.

[122] Tomás J, Mas JÀ, Casacuberta F. A quantitative method for machine translation evaluation. In: Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable? 2003 Apr 12–17; Budapest, Hungary; 2003.

[123] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; 2005 Jun 29; Ann Arbor, MI, USA; 2005.

[124] Tsiartas A, Georgiou PG, Narayanan SS. Toward transfer of acoustic cues of emphasis across languages. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association; 2013 Aug 25–29; Lyon, France; 2013.

[125] Do QT, Sakti S, Nakamura S. Sequence-to-sequence models for emphasis speech translation. IEEE/ACM Trans Audio Speech Lang Process 2018;26(10): 1873–83.

[126] Do QT, Toda T, Neubig G, Sakti S, Nakamura S. Preserving word-level emphasis in speech-to-speech translation. IEEE/ACM Trans Audio Speech Lang Process 2017;25(3):544–56.