文章编号：1003-0077(2019)11-0001-14

# 神经网络机器翻译研究热点与前沿趋势分析

林 倩，刘 庆，苏劲松，林 欢，杨 静，罗 斌

（厦门大学 信息学院，福建 厦门 361005）

**摘 要**：机器翻译是指利用计算机将一种语言文本转换成具有相同语义的另一种语言文本的过程。它是人工智能领域的一项重要研究课题。近年来，随着深度学习研究和应用的快速发展，神经网络机器翻译成为机器翻译领域的重要发展方向。该文首先简要介绍近一年神经网络机器翻译在学术界和产业界的影响，然后对当前的神经网络机器翻译的研究进展进行分类综述，最后对后续的发展趋势进行展望。

**关键词**：人工智能；深度学习；神经网络机器翻译

**中图分类号**：TP391 **文献标识码**：A

## Focuses and Frontiers Tendency in Neural Machine Translation Research

LIN Qian, LIU Qing, SU Jinsong, LIN Huan, YANG Jing, LUO Bin

(School of Informatics, Xiamen University, Xiamen, Fujian 361005, China)

**Abstract**: Machine translation is the process of to attempting convert text from one language to another using computers, which has already become the research issues of great importance in artificial intelligence. With the fast growth of deep learning research and application, it has been revealed that neural machine translation become a mainstream of research for machine translation. This paper firstly introduces the influence of neural machine translation in academia and industry in the past year, and then reviews the research progress on neural machine translation, finally we outline the outlook for its future development.

**Keywords**: artificial intelligence; deep learning; neural machine translation

## 0 引言

机器翻译是利用计算机自动将一种语言翻译成另外一种语言的过程，它是人类长久以来的一个梦想。随着全球化进程的加速和互联网的快速发展，各国之间的信息交流日趋频繁，高效、快速的翻译逐渐成为人们的日常需求之一。然而，传统人工作业的翻译方式无法满足迅猛增长的翻译需求。而机器对海量数据的处理能力使得高效、快速的翻译成为可能，人们的目光开始转向机器翻译，对于机器翻译的需求空前增长。

近年来，随着深度学习的快速发展，神经网络机器翻译研究取得了巨大进展。在许多语种上，神经网络机器翻译（neural machine translation，NMT）性能大幅度提升，远远超过了传统统计机器翻译（statistical machine translation，SMT）。目前，NMT 研究是自然语言处理研究的热门前沿发展方向。表 1 给出了 2017 年和 2018 年人工智能、自然语言处理方向的各大权威期刊和会议与 NMT 主题相关的 Regular 论文统计情况。发表论文的数量充分表现出 NMT 研究已经成为学者们关注的焦点，在学术界有着巨大影响力。

与此同时，产业界各大公司也投入人力、物力研发各自的神经网络机器翻译实用系统。NMT 已成为各大机器翻译系统的核心技术，促进了翻译工具市场的蓬勃发展。翻译工具日益普遍，已经成为人们生活中的重要组成部分，从文献翻译到国际交流，

收稿日期：2018-12-28 定稿日期：2019-04-26

基金项目：国家自然科学基金(61672440)；国家语委一般项目课题(YB135-49)；厦门大学校长基金(ZK1024)；国家重点研发计划(2019QY1803)

机器翻译正在发挥重要作用。可以说,机器翻译的发展在学术界和产业界都已经完全进入了NMT时代。

纵观近一年发表的NMT论文,我们可以粗略得知该领域的研究发展主要集中在以下几个方向:词汇表受限研究、资源受限研究和模型研究,而模型研究是整个NMT研究的重点。接下来,本文将对这三个方面分别进行介绍。最后,对NMT的后续发展进行了展望。

表1 2017—2018年神经网络机器翻译的Regular论文发表篇数

| 会议/期刊 | 2017年 | 2018年 |
|---|---|---|
| ACL | 14 | 12 |
| EMNLP | 15 | 22 |
| NAACL | 0 | 14 |
| COLING | 0 | 18 |
| TACL | 4 | 3 |
| TASLP | 2 | 5 |
| AAAI | 3 | 10 |
| IJCAI | 4 | 2 |
| ICLR | 3 | 6 |
| NIPS | 3 | 0 |
| ICML | 1 | 1 |
| 总计/篇 | 49 | 93 |

## 1 词汇表受限研究

传统NMT模型使用固定大小的词表,编码器无法学习词表之外(out-of-vocabulary,OOV)的词的语义表示,解码器无法选择词表之外的词作为译文,极大地影响了翻译质量。因此,如何解决词汇表受限问题已成为研究的重点之一。当前的研究主要通过基于细粒度语义单元的建模方法来解决该问题。这方面的研究主要分为以下三类:

(1)字词混合的语义建模方法。Passban等[1]用字符级的解码器提高了形态学丰富的语言翻译质量;Chen等[2]在编码器中使用了词和字符两个粒度的信息,在解码端用多个注意力,使不同粒度的信息能够协同帮助翻译;Zhao等[3]为建模同一语系的语言对之间的相似性,其编码器由字符级单向RNN和词级双向RNN组成,并使用自顶向下的层次注意力机制,先获得词级别的上下文,再获得字符级的上下文,共同用于预测目标语言字符。

(2)基于子词的语义建模方法。这类工作主要是以BPE(byte pair encoding)[4]为基础。Kudo等[5]提出子词正则化,利用一元语言模型生成多种候选的子词序列,丰富NMT编码器的输入以增强翻译系统的鲁棒性。Morishita等[6]引入了多粒度BPE的表示来平均求得词汇语义表示。特别地,Morishita等[6]认为编码器词向量层、解码器词向量层以及解码器输出层有着不同的作用,因此不同层的BPE粒度的选择也应该有所差别。

(3)词干加词缀的语义建模方法。Song等[7]将词分解为词干和后缀,解码时先生成词干,再生成后缀;为了解决形态丰富的语言中存在源语言单词可能对应多个目标语言单词的情况,Passban等[8]将源语言词拆分成词干和后缀作为编码器的双通道输入,相应地,在解码器端使用了双注意力机制来捕获编码器不同粒度的语义信息。

这三类方法都在一定程度上缓解了词表之外的词降低翻译质量的问题。字词混合的语义建模方法结合了词级方法和字符级方法两者的优点,既可以缓解以词为语义单元造成的词汇表受限问题,又避免了完全只使用字符信息造成语义单元歧义大、输入序列过长的问题。词缀等语言现象的存在使得基于子词的建模方式受到了越来越多的关注。词干加词缀的建模方法则为形态学丰富语言的翻译任务提供了新思路。

## 2 资源受限NMT研究

NMT模型以平行句对为基础。然而,对于许多语言的翻译任务而言,平行句对的获取并不容易。因此,如何在资源受限的情况下建立高性能的NMT模型也成为研究热点之一。

### 2.1 无监督NMT研究

无监督NMT致力于在只有单语数据的情况下构建翻译模型。

Lample等[9]总结了无监督机器翻译取得成功的三个重要步骤:初始化、语言模型和迭代的反向翻译;Lample等[10]不依赖任何平行语料,只使用单语数据集来进行翻译模型建模。具体实现中,Lample等[10]利用降噪自编码器和对抗训练将两种语言映射到相同的隐式空间,并迭代训练两个方向的翻

译模型；Artetxe 等[11]先预训练词向量，利用自编码器和反向翻译实现无监督 NMT；Yang 等[12]认为之前的无监督 NMT 使用共享编码器来编码不同语言的语义表示容易丢失不同语言各自的特性，进而限制翻译性能。对此，Yang 等[12]提出每种语言应该使用各自的编码器进行建模，只对编码器的后几层和解码器的前几层的权重进行共享。

### 2.2　半监督 NMT 研究

与传统 NMT 和无监督 NMT 不同，半监督 NMT 以大量单语语料和少量平行语料为基础来进行翻译模型建模。

Fadaee[13]首先分析发现：反向翻译生成的伪平行数据太多、容易使得 NMT 模型更偏向其中的噪声数据。此外，伪平行数据的使用对于具有高预测损失的单词最有帮助。进一步地，Fadaee[13]提出先识别目标语言中难以预测的单词，然后在单语数据中对含有这些词的句子进行采样，增加这些单词的出现次数，并在单语数据中对类似于难预测词上下文的句子进行采样；但是另一方面，反向翻译的模型性能仍受限于合成语料的质量。因此 Zhang 等[14]提出了联合训练方法，对源语言到目标语言和目标语言到源语言的 NMT 模型进行联合训练，一方向的 NMT 模型为反方向的 NMT 模型提供伪平行数据，如此迭代多次，可以同时提升 NMT 模型的翻译性能；基于同样的想法，Wang 等[15]采用对偶学习来进行两个方向的半监督 NMT 联合建模。

### 2.3　基于枢轴的 NMT 研究

基于枢轴的 SMT 研究取得了很好的效果。研究者将这个想法迁移到 NMT 中，也取得了不错的效果。

Chen 等[16]通过两种语言的多模态数据实现零资源翻译，首先利用源语言的文本—图片数据，训练源语言的图片描述（image captioning）模型，用于生成目标语言图片对应的源语言描述，以此来构造伪平行文本数据训练翻译模型；Ren 等[17]提出在含有稀缺资源的语言对中引入高资源语言，将低资源的语言作为中间隐变量，以最大化大规模平行语料的似然为目标，使用双向 EM 算法联合训练两个方向共四个 NMT 模型。

### 2.4　领域自适应研究

与 SMT 研究一样，领域自适应也一直是 NMT 研究的重点。让非目标领域的平行语料来帮助建立更好的目标领域翻译模型，可以在一定程度上解决资源受限的问题。

从语料的角度，Zhang 等[18]计算目标领域和候选训练平行句对的语义相似度，然后将相似度融入目标函数来实现模型领域自适应；Wang 等[19]为解决领域自适应，提出了两种方法：句子选择和句子加权，并能够在训练过程中动态进行句子选择和权重计算。上述工作均是从语料选择的角度来解决领域自适应问题。从建立模型的角度，Zeng 等[20]用多任务学习方法联合训练神经网络机器翻译模型和基于注意力机制的领域分类器，分别学习到领域相关和领域无关的上下文信息，并且通过强化领域相关的目标语言词来优化模型训练。从参数生成的角度，Ha 等[21]和 Platanios 等[22]均致力于如何使用小的网络根据上下文信息来为每个翻译句子动态生成模型参数。特别地，Chu 等[23]对 NMT 中的领域自适应研究进行了归类总结。

### 2.5　多模态 NMT 研究

近年来，融合文本之外的其他模态信息成为了 NLP 的一个研究热点。同样，在机器翻译领域，融合了文本、图像等模态信息的多模态 NMT 研究也成为了 NMT 发展的新趋势。

与传统的多模态 NMT 模型不同，Delbrouck 等[24]在使用 CNN 抽取图片信息时，充分考虑文本信息，同时使用注意力机制捕获图片的相关信息来作为文本语义表示的有效补充；而 Zhou 等[25]则是引入多任务学习，同时建模两个任务：多模态翻译，图片—文本的联合语义表示。

### 2.6　多语言多任务 NMT 研究

多语言多任务 NMT 一直是 NMT 研究的热点，它的优势在于可以充分发挥神经网络模型语义表示向量化、参数共享的优势。

在这方面，Gu 等[26]通过多种语言之间的词汇和句子语义表示共享，使得低资源语言 NMT 模型能够利用高资源语言 NMT 模型的词汇和句子表示；Blackwood 等[27]针对多语言 NMT 提出了一个任务相关的注意力模型，对不同目标语言采用不同的注意力参数，提高了多语言 NMT 模型的性能。Gu 等[28]引入元学习（meta learning）来实现从多种资源丰富语言对的翻译模型到低资源语言对翻译模型的模型参数快速自适应；Wang 等[29]为多语言

NMT提出三个改进策略：在解码器中使用初始化标签引导目标语言翻译、引入位置信息，并将隐状态分割为语言共享和语言独立的两个单元。多任务机制同样是解决资源受限的一个方法。Kiperwasser等[30]认为，训练时多任务中的辅助任务权重应该先大后小，而主要任务反之，并在此基础上提出了根据训练时间变化来动态决定每个任务在不同训练阶段的权重；Zaremoodi等[31]通过多任务学习对机器翻译任务和语义分析、语法分析和命名实体识别等辅助任务进行联合建模，使得NMT模型能够自动学习到语义和句法知识；Niu等[32]在多任务学习框架下考虑了和机器翻译相关的两个任务：单语正式度转换(formality transfer)和正式度敏感的机器翻译(formality sensitive machine translation)。

## 3 模型研究

### 3.1 模型架构

从模型框架的角度来进行分类，目前NMT模型主要包含三类：基于循环神经网络的NMT模型(recurrent neural network based NMT，RNMT)，基于卷积神经网络的NMT模型(convolutional sequence to sequence learning，ConvS2S)，基于自注意力机制的NMT模型(Transformer)。下面我们对这三类模型一年来的进展进行介绍。

#### 3.1.1 RNMT

如图1所示，RNMT模型主要包含基于双向循环神经网络(recurrent neural network，RNN)的编码器，基于单向RNN的解码器两个部分，其主要特点是在解码时每步都使用注意力机制动态地捕获与当前译文相关的源语言上下文信息。针对RNMT模型，研究人员在原本的网络结构基础上做了大量的尝试和改进，包括编码器和解码器的改进、信息建模方式的创新及外部知识的引入等。

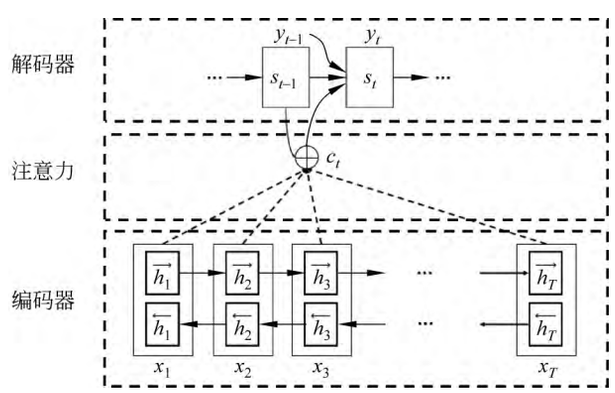


图1 RNMT框架图

编码器 NMT的编码器以RNN为基础。由于RNN存在梯度消失和爆炸的缺陷，因此无法很好地建模长距离信息。对此，Su等[33]提出了将输入句子进行切分，形成词—子句—句子的层次结构，然后引入层次循环神经网络来进行编码器建模。同时，解码器以子句为单元进行逐子句翻译，并且引入两个注意力机制来分别捕获子句内部和子句之间的上下文信息。Zhang等[34]在编码层增加了一个关系网络层，该层可以有效建模不同源语言词对应的语义表示(annotation)之间的关系，改善编码器的语义建模效果。此外，现有RNMT编码器只使用双向RNN得到的annotation来表示输入句子的语义，存在信息表示不充分的缺陷。针对该缺陷，Xiong等[35]设计了一种多信道编码器，包含来自输入源语言词的向量表示、双向RNN的隐藏状态、神经图灵机中的外部存储，并进一步引入门机制来自动学习不同信道语义表示的权重。受计算机视觉稠密卷积神经网络的启发，Shen等[36]也对RNMT的编码器和解码器进行修改，使得当前隐状态的语义表示生成同时受前面所有隐状态的影响，既在一定程度上解决梯度消失爆炸问题，又增强了模型的语义表示能力。

注意力机制 注意力机制主要是使用目标隐状态来对源语言annotation进行相关度权重计算(本质是计算能量函数再进行归一化)，自动生成和当前相关译文选择的上下文信息。Werlen等[37]利用注意力机制将之前预测的所有词信息融入当前词预测过程。Wang等[38]也用同样的思想将注意力机制的关注范围扩展到目标语言的隐状态，在中英实验上获得了最佳性能。受到深层编码器和解码器工作的启发，Zhang等[39]将注意力机制也改为深层模型，使得每一层的编码器都能获得相应层次级别的上下文信息。由于全局注意力机制会将权重分散到所有的源语言词汇，削弱了关键性词汇的影响。为了能够将注意力集中在关键信息上，研究者们提出了局部注意力机制[40]只对序列窗口内的内容进行建模，然而局部注意力通常以对齐词为中心，权重往两边递减，这样的做法不一定合理：一是按照绝对距离来衡量权重不合理；二是上下文窗口无法保证包含所有重要的上下文词。Chen等[41]在局部注意力机制的基础上进一步加入语法信息约束，在依存树上

设置窗口，对依存树上中心词周围的词进行关注建模，这样能够捕获绝对距离较远但语法距离较近的词的信息，从而获得更好的上下文表示。为了能够利用更多的信息，许多工作引入记忆(memory)模块来记录历史注意信息，辅助网络学习。记忆模块主要承担两个功能：记忆注意力机制历史信息和更新句子表示。在此基础上，Meng 等[42]提出使用两个记忆模块来分别承担两个功能，从而进一步改善记忆模块的使用效果。

解码器　在传统的 RNMT 中，解码器利用三部分信息(由注意力机制捕捉到的源语言上下文、前一时间步的隐状态和前一时间步的译文)来预测当前时间步的译文。在此基础上，Li 等[43]提出首先预测目标译文的词性，预测得到的词性可以用于帮助注意力机制生成更好的源语言上下文，以改善最终的译文预测。Huang 等[44]提出了基于短语的 NMT 模型，该模型能够显式地建模输出序列中的短语结构，作者利用“sleep-wake network”[45]来对齐目标语句与源语句，代替传统的注意力机制。

现有 NMT 模型在解码时的 Softmax 操作十分耗费时间，为了解决该问题，Shi 等[46]借助词对齐信息，减少目标语言的候选词汇，从而加快解码速度。借鉴传统 SMT 解码工作的成功经验，Zhang 等[47]提出面向 NMT 的立方体剪枝方法，其主要思想是通过合并前缀译文相似的译文假设来构造等价类，每个类进行各自的 Softmax 译文选择操作。搜索时，挑选所有类中分数最小的译文假设进行扩展搜索。这种方式不仅能减少解码器的搜索次数，同时也减少了 Softmax 操作。与之前的工作不同，Post 等[48]主要是探索如何在具有候选译文约束的情况下进行译文搜索，论文作者提出了新的搜索算法，使得搜索时间复杂度与候选译文约束的个数无关。

其他网络结构改进　传统的 RNMT 模型存在一个明显的缺陷：解码器输出与编码器输入缺乏直接联系，因此在模型后向求导过程中容易出现梯度消失和梯度爆炸的问题。对此，Kuang 等[49]通过增加解码器目标译文和源语言输入词汇之间的联系来解决该问题。

针对 RNMT 低频词存在训练不足的难题，Nguyen 等[50]进行了两种改进：①在标准的输出层计算中加入正则化操作；②引入一个简单的词汇模块来解决原本给予高频词过多反馈的问题，从而改善低频词的翻译情况。与前面工作不同，Liu 等[51]则是利用上下文信息来优化源语言输入词的嵌入表示，进而优化译文选择。而 Wang 等[52]则是针对 NMT 翻译过程中代词缺失现象进行研究，提出引入重建网络来使得编码器和解码器隐状态能够重建代词信息，强化模型翻译代词的能力。

此外，还有一些工作主要致力于对 RNMT 神经网络单元的改进。例如，Li 等[53]在原有的权重矩阵外再乘上一个由神经网络计算出的动态权重，进一步动态区分了神经网络单元中加权求和操作中不同部分的作用。Zhang 等[54]则使用简单的加减操作简化门控循环单元，只保留了权重矩阵。这样的建模方式加快了计算速度，并且使得隐层状态具有可解释性。

模型压缩　现有 NMT 模型往往模型结构复杂，参数量巨大。因而，模型压缩也是一种 NMT 研究选择。剪枝是实现模型压缩的方法之一。See 等[55]采用了参数剪枝，以很小的性能代价压缩模型，解决了深度模型参数量过大的问题。Shu 等[56]用一组编码表示词汇，压缩了词嵌入表示，大大减少了模型在词嵌入表示部分的参数。

未来信息建模　与上述工作不同，近期还有许多研究工作涉及到了 NMT 的多次解码建模。这类工作的思想与人工翻译往往需要多次修改的过程是一致的。例如，Xia 等[57]引入两次解码来优化译文生成，其中第二次解码时会参考第一次解码的译文信息。而 Zhang 等[58]主要是考虑了反向和正向译文的互补性，提出了引入反向解码器进行反向解码产生反向译文信息，然后再进行正向解码。在这过程中，正向解码器同时关注编码器和反向译文信息，因此能够生成更好的译文。在前面工作基础上，Geng 等[59]则是引入增强学习，根据输入句子的翻译难度和已产生译文的质量来自动决定多次解码的次数。Su 等[60]和 Schulz 等[61]都致力于引入变分循环神经网络来增强 NMT 译文的多样性。Lin 等[62]则是使用编码器利用反卷积操作产生全局语义信息，再利用注意力机制将其融入解码器。而 Zheng 等[63]则是通过在解码器中引入两层网络分别建模已生成译文和未生成译文，并同时建模了二者的语义关系。显然，后者信息可用于优化译文生成。

跨句子信息建模　传统 NMT 翻译都是逐句进行翻译。然而，人们在翻译过程中，往往会用到跨句子的信息，这样翻译出来的译文才会更加完整连贯。基于此，Kuang 等[64]通过引入门机制来动态控制前一个句子有多少信息被用于当前句子的翻译过程

中。Tu等[65]提出使用缓存(cache)来保留先前(跨句子和翻译句子当前状态之前的)的上下文和译文信息,然后在解码过程中使解码器根据上下文的相似性从cache中读取之前的隐状态,优化当前时间步的隐状态建模表示。Maruf等[66]引入两个记忆模块来分别存储文档中源语言句子信息和第一次翻译产生的目标句子信息,这样在第二次翻译时通过关注这两个模块,NMT模型就可以充分利用源语言和目标语言文档级别的上下文信息。Kuang等[67]则是采用两个缓存来分别存储目标语言句子级别动态产生的主题词和文档级别静态产生的主题词,这种方式可以方便解码器使用文档级别的信息。

**SMT知识** SMT模型架构完全不同于NMT。学到的翻译知识也有别于NMT学到的翻译知识。因此,引入SMT的翻译知识来改进NMT模型也是一种研究选择。

Wang等[68]在译文选择时不仅考虑了神经网络的预测概率,还同时考虑SMT的译文预测概率。神经网络机器翻译系统可能存在生成文本很流利,但是翻译得不够准确的问题,Zhao等[69]构造了基于目标语言的前缀树来存储双语短语和对应的翻译概率,使得解码器挑选译文时能尽量使用前缀树包含的译文。此外,Zhao等[70]使用记忆模块来存储低频词的上下文信息和译文嵌入表示,用于辅助低频词的后续翻译。

**句法知识** 基于句法的SMT研究的成功证明了句法知识对翻译的重要性。因而,利用句法知识来改进NMT模型也成为了研究者的选择。这方面的前沿工作包括:Chen等[71]将基于句法树的RNN扩展为双向,改进了编码器建模。同时,在解码器端引入了针对树结构的覆盖度模型(tree-coverage model)。Li等[72]先将源语言句子句法信息进行序列化,然后再用不同方法进行RNN建模,以此来将源端句法信息融入编码器。而Wu等[73]则是研究如何利用目标译文的句法信息来改善译文生成。具体而言,他们提出在生成目标端译文的同时构建译文的依存结构,并将依存信息用于辅助下一个目标词的生成。Wu等[74]提出了编码器和解码器都是基于依存树的NMT模型。为了克服基于1-best句法树在NMT建模上存在的缺陷,Ma等[75]和Zaremoodi等[76]着重研究如何基于短语树森林来进行NMT建模。前者主要关注如何将短语树森林进行序列化,以方便后续的双向RNN建模,而后者主要关注如何直接基于短语树森林进行自底向上的语义融合生成。与前面工作的解码方式存在明显不同,Gu等[77]提出了一种自顶向下带有句法信息的译文生成方式,以此来充分利用目标语言译文的句法信息。另外,Bastings等[78]引入图卷积网络(GCN),在传统编码器—解码器结构翻译模型的编码器端加入图卷积层来引入句法依存树信息,作者分别尝试了CNN、BOW和双向RNN三种类型的编码器,实验结果表明三种模型与GCN结合后效果均有提升。

**其他外部知识** 除了上述信息,研究者们还探索了其他类型外部知识对NMT模型翻译效果的影响。例如,Li等[79]在编码端加入了一个知识模块,该模块存储了额外的语言学信息,把这些信息作为输入词汇的补充,使得编码器能够包含尽可能多的语言学信息。Gu等[80]和Zhang等[81]则是利用搜索引擎检索得到的信息来改善翻译质量。而Ugawa等[82]则是在编码器中引入了命名实体信息。

### 3.1.2 Transformer

传统NMT模型多是基于双向RNN进行序列化建模,即当前时刻的隐状态语义表示只直接依赖于上一时刻的隐状态语义表示和当前时刻的输入信息。Vaswani等[83]提出的Transformer框架引入了多重自注意力(multi-head self-attention)机制来增强模型翻译能力。

与RNN相比,自注意力机制中当前节点的隐状态语义表示同时依赖于序列中所有节点的隐状态表示,具有更强语义建模能力和可并行化训练的优点。如图2所示,编码器端首先对输入的源语言词的嵌入表示进行多重自注意力机制建模,获得源语言句子的最终语义表示。略有不同的是,解码器引入Masked多重自注意力机制建模目标语言的上下文信息,这个信息再和编码器的语义表示进行多重注意力机制建模,生成源语言的上下文信息,最后得出概率分布。因此,Transformer很快成为机器翻译界的新宠儿,如何基于Transformer进行改进成为了NMT研究的热点问题。

标准的Transformer以序列生成的方式来产生译文,因此,翻译模型在测试阶段无法实现译文的并行生成,影响了模型效率。对此,Gu等[84]提出首先预测源语言词的繁殖度(fertility),然后根据繁殖度分别进行复制,复制后的源语言词同时输入解码器,以此来实现目标译文的并行生成。这种方式虽然解决了标准模型无法并行的问题,但也丢失了目标译文之间的序列信息。对此,Wang等[85]则提出了一

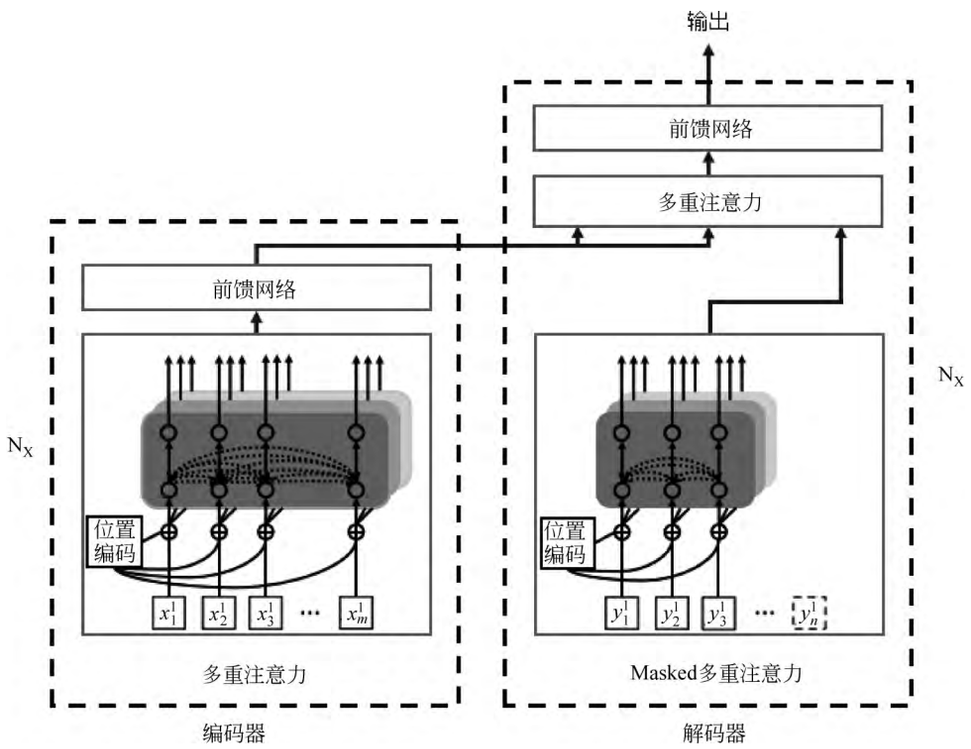


**图 2　Transformer 框架图**

种折中方案：目标译文分组生成，组内的译文并行生成，组间保留顺序生成。与上述关注如何并行生成译文的工作不同，更多工作都致力于 Transformer 中的自注意力机制的改进研究。目前的研究主要分两类：一类做法是加入额外信息来优化注意权重计算。例如，Shaw 等[86]引入了新的位置矩阵以建模词之间的相对位置关系；Yang 等[87]用自注意力的中间状态计算得到一个高斯偏置项，并将其加入到原来的自注意力分布计算中，使权重分布更为平滑，以提升捕获短距离语义依赖的能力；另一类做法是改进自注意力机制的权重计算方式。例如，Shen 等[88]提出了具有方向性、多维度的自注意力机制；Zhang 等[89]将解码器端的 Masked 自注意力机制替换成平均注意力机制，在保证模型效果可比的情况下加快了模型的解码速度；Li 等[90]引入子空间不一致、注意位置不一致和输出表示不一致三种正则化项，分别促进了子空间、注意位置和输出表示的差异性，来确保多重自注意力机制的多样性和互补性；Shen 等[91]提出了一种双向分块自注意力机制，将输入的词嵌入序列划分为等长的分块，先进行块内的自注意力计算，再对上一步结果进行块间的自注意力计算，实现更快且节省空间的上下文融合。

此外，还有不少研究者关注如何对层与层之间的信息进行融合，以此来减少编码器信息丢失，提升模型效果。Dou 等[92]提出了两种融合方法：层回归和多层注意力。其中层回归把每一层同一个位置的隐状态通过残差连接、线性组合、迭代组合或层次组合的方式融合，而多层注意力机制将计算自注意力机制时的加权对象由当前层扩展到了当前层以下的每一层，最后将每层各个位置的隐状态都进行融合。Wang 等[93]则是对每层同一位置的隐状态语义表示进行了融合。

3.1.3　ConvS2S

卷积神经网络(convolutional neural network，CNN)通常用于图像信息抽取，卷积操作能够并行计算，提升模型效率。为了避免 NMT 序列化建模的缺陷，基于 CNN 来进行 NMT 建模也是选择之一。

如图 3 所示，Gehring 等[94]提出 ConvS2S 模型，将 CNN 引入序列到序列的翻译模型中，编码器和解码器采用相同的卷积操作，然后经门控线性单元进行非线性变换得到相应输出。值得关注的是，注意力机制为多跳注意力，即每个卷积层都进行注意力建模，上一层的卷积的输出作为下一层卷积的输入，经过堆叠得到最终的输出。而 Gehring 等[95]用卷积神经网络建模了 NMT 编码器。Kaiser 等[96]将在图像分类任务中取得很好效果的深度可

分离卷积网络应用于神经网络机器翻译，减少了卷积操作中的参数数量。以上几个工作均在保证一定翻译准确度的情况下提升了翻译速度。

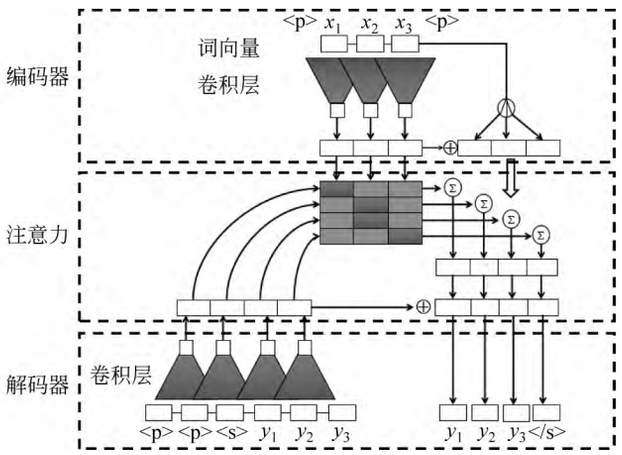


图3 ConvS2S框架图

### 3.2 模型对比和分析

纵观NMT的发展历史，不同模型框架有着不同的优势和缺陷。自然地，对这些模型进行对比和分析也成为NMT研究领域的一个研究方向。

Chen等[97]将当前效果显著的几种优化算法与模型结合，探究不同优化算法对不同模型的影响。得出如下结论：①标签平滑[98]对RNMT和Transformer都有效；②多重注意力机制对RNMT和Transformer都有效；③层标准化使得RNMT和Transformer的模型训练更加稳定；④训练时增大Batch Size大小对RNMT和Transformer都有效。Domhan等[99]的研究表明通过调节参数和增加优化算法，RNMT、ConvS2S的性能能够达到和Transformer可比的程度。

通过对翻译模型本身各部分的重要性进行粒度分析，Domhan等[99]认为编码端最后一层语义表示对于翻译模型最为重要。此外，多重注意力机制和残差前馈层的作用也不容忽视，而其中源语言的自注意力机制比目标语言的自注意力机制更为重要。Wang等[100]把传统注意力机制的作用范围扩展到解码器隐状态，使得RNMT的翻译性能第一次超过了Transformer。Lakew等[101]针对多语言翻译进行了三个方面的对比：一是对双语、多语、零样本系统的翻译质量进行定量比较；二是对RNMT和Transformer的翻译质量进行比较；三是考察语言接近程度对于零样本翻译的影响。最终论文得出如下结论：①多语言模型的性能要比双语模型性能更强；②利用多语言语料训练，Transformer相对于RNMT的性能提升更加明显；③使用相关语言的语料能有效提升多语言模型的性能；④在双语和零样本模型中，源语言和目标语言的相关性对模型的性能影响不大；⑤对于零样本的情况，Transformer模型比RNMT表现得更好。

Tang等[102]同样对比了RNN，自注意力机制和CNN在两个与翻译密切相关任务上的性能：主谓一致和词义消歧。论文发现在主谓一致任务的长距离建模上CNN和自注意力机制的模型效果并不会优于RNN，只有当建模距离长于一定值时两者的效果才会与后者相当，甚至更好。

除了上述研究，还有一些工作侧重于探究模型性能受限的原因，以及解决特定问题的能力。Ott等[103]提出当前NMT模型性能受到答案的多样性和噪声训练数据的限制。

Belinkov等[104]在基于字符的NMT模型上将字符嵌入表示的平均来作为词嵌入表示，使得模型不受输入字符顺序错位的影响。此外，还采用了对抗实例集成训练的方法，使得模型能够同步学习对多种噪声具有鲁棒性的语义表示。Tan等[105]则对Transformer模型进行了语言学分析，认为语言学特性对模型正确率的影响比错误传播更大。

### 3.3 模型训练

对NMT模型训练方法的改进研究主要集中在增强学习和对抗学习两个领域。增强学习通过对机器当前的每一步行为给予不同奖励，来指导模型自动选择如何做出正确的决策。通常，奖励函数设置为和译文评价指标直接相关的函数，因而能够在一定程度上解决NMT模型训练和测试评价函数不一致的问题。由于在增强学习中往往需要对译文进行采样生成，使用基于对数似然的训练方法面临着训练和测试不一致的问题，即模型在测试时必须根据前面做出的决策来生成标记，而无法像训练阶段一样使用正确标记信息。Bahdanau等[106]和He等[107]分别引入判定网络和价值网络来估计采样生成译文的质量。Wu等[108]针对增强学习在深层模型和大规模数据训练上的不稳定缺陷，提出设置更有效的奖励函数，将原来目标函数加入到增强学习目标函数中，以使得增强学习训练更加稳定。

对抗学习的核心思想是让两个目标相反的网络交替训练，这两个网络分别被称作生成器和判别器。生成器的目的是产生判别器无法区分真假的数据，

而判别器要尽量将真假数据分辨出来，两者相互抗衡可以使得模型整体上达到更好的效果。生成器和判别器都可看作是黑盒模型。Ebrabimi 等[109]不再将生成器当作黑盒，而是提出几种修改原文本的方式，使模型学会生成改变方式，并且从多种可能的句子中选择对抗对象。这样做的好处在于可以提升对抗样本的质量，覆盖更多种的对抗方式，有利于增强模型的鲁棒性。Cheng 等[110]则是将编码器作为生成器，带有噪声的源语言句子和原本的源语言句子同时经过编码器语义建模，训练判别器将两者语义表示区分开来。传统的对抗学习中生成器往往不可导，Gu 等[111]引入耿贝尔分布(gumbel distribution)使得生成器可导，模型可以实现一体化训练。Yang 等[112]结合了增强学习和对抗学习的思想，通过判别器给出的相似度作为增强学习的奖励函数，以此来进行判别器与生成器的交替训练。

## 4　总结和展望

综上所述，神经网络机器翻译技术正在发挥越来越重要的作用，在学术界和产业界有着巨大的影响力，已经成为机器翻译领域的主流技术。但是，神经网络机器翻译仍然面临诸多挑战，未来的发展趋势值得更多的关注。

(1) 资源受限的 NMT 研究。资源问题一直是困扰 NMT 研究和产业化的首要问题。随着 NMT 产业化的逐渐推广，这个问题将日益突出。因此资源问题仍将是本领域研究的重要问题。

(2) 知识驱动的 NMT 研究。人工翻译融合了多方面、多维度的知识。因此，要构建一个高性能的 NMT 模型，如何融合除平行语料之外的翻译知识也是进一步提升 NMT 模型效果的关键所在。

(3) NMT 模型简化研究。目前，基于 RNMT+[97]和 Transformer 都取得了非常好的翻译效果。然而，随着翻译性能不断提升，带来的问题是模型变得日益复杂。如何在保持翻译性能的前提下，对这些翻译模型进行简化，降低训练复杂度，将是 NMT 产业化过程中需要解决的问题。

(4) NMT 模型可解释性研究。神经网络的可解释性一直是深度学习的研究重点，相较于计算机视觉、图像处理，基于神经网络的自然语言处理在模型可解释性方面的研究更为缺乏，NMT 研究也不例外，模型可解释性研究将有助于我们进一步推动 NMT 其他方面研究的进展。

(5) 新 NMT 架构设计。目前 NMT 架构主要以 RNMT、Transformer 和 ConvS2S 为主。三类模型架构性能相当，各有优缺点。如何融合三类架构的优点，设计出性能更好的翻译架构，也是学术界不断探索的研究问题。

## 参考文献


[1] Passban P, Liu Q, Way A. Improving character-based decoding using target-side morphological information for neural machine translation[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics, 2018: 58-68.

[2] Chen H, Huang S, Chiang D, et al. Combining character and word information in neural machine translation using a multi-level attention[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics, 2018: 1284-1293.

[3] Zhao S, Zhang Z. Attention-via-attention neural machine translation [C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 563-570.

[4] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2016: 1715-1725.

[5] Kudo T. Subword regularization: Improving neural network translation models with multiple subword candidates[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2018: 66-75.

[6] Morishita M, Suzuki J, Nagata M. Improving neural machine translation by incorporating hierarchical subword yeatures[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 618-629.

[7] Song K, Zhang Y, Zhang M, et al. Improved English to Russian translation by neural suffix prediction[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 410-417.

[8] Passban P, Way A, Liu Q. Tailoring neural architectures for translating from morphologically rich languages[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 3134-3145.

[9] Lample G, Ott M, Conneau A, et al. Phrase-based & neural unsupervised machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 5039-5049.

[10] Lample G, Conneau A, Denoyer L, et al. Unsupervised

machine translation using monolingual corpora only [C]//Proceedings of the International Conference on Learning Representations,2018: 1-14.

[11] Artetxe M, Labaka G, Agirre E, et al. Unsupervised neural machine translation [C]//Proceedings of the International Conference on Learning Representations,2018: 1-12.

[12] Yang Z, Chen W, Wang F, et al. Unsupervised neural machine translation with weight sharing [C]//Proceedings of Meeting of the Association for Computational Linguistics,2018: 46-55.

[13] Fadaee M, Monz C. Back-translation sampling by targeting difficult words in neural machine translation [C]//Proceedings of the Empirical Methods in Natural Language Processing,2018: 436-446.

[14] Zhang Z, Liu S, Li M, et al. Joint training for neural machine translation models with monolingual data [C]//Proceedings of the National Conference on Artificial Intelligence,2018: 555-562.

[15] Wang Y, Xia Y, Zhao L, et al. Dual transfer learning for neural machine translation with marginal distribution regularization [C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 5553-5560.

[16] Chen Y, Liu Y, Li V O. Zero-Resource neural machine translation with multi-agent communication game[C]//Proceedings of the National Conference on Artificial Intelligence,2018: 5086-5093.

[17] Ren S, Chen W, Liu S, et al. Triangular architecture for rare language Ttranslation [C]//Proceedings of the Meeting of the Association for Computational Linguistics,2018: 56-65.

[18] Zhang S, Xiong D. Sentence weighting for neural machine translation domain adaptation[C]//Proceedings of the International Conference on Computational Linguistics,2018: 3181-3190.

[19] Wang R, Utiyama M, Finch A M, et al. Sentence selection and weighting for neural machine translation domain adaptation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2018, 26 (10): 1727-1741.

[20] Zeng J, Su J, Wen H, et al. Multi-domain neural machine translation with word-level domain context discrimination[C]//Proceedings of the Empirical Methods in Natural Language Processing,2018: 447-457.

[21] Ha D, Dai A M, Le Q V, et al. Hyper Networks[C]//Proceedings of the International Conference on Learning Representations,2017: 1-18.

[22] Platanios E A, Sachan M, Neubig G, et al. Contextual parameter generation for universal neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing,2018: 425-435.

[23] Chu C, Wang R. A survey of domain adaptation for neural machine translation [C]//Proceedings of the International Conference on Computational Linguistics,2018: 1304-1319.

[24] Delbrouck J B, Dupont S. Modulating and attending the source image during encoding improves Multimodal Translation [C]//Proceedings of the NIPS 2017 Workshop on Visually-Grounded Interaction and Language (ViGIL),2017: 1-9.

[25] Zhou M, Cheng R, Lee Y, et al. A visual attention grounding neural model for multimodal machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 3643-3653.

[26] Gu J, Hassan H, Devlin J, et al. Universal neural machine translation for extremely Low resource languages[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics, 2018: 344-354.

[27] Blackwood G W, Ballesteros M, Ward T. Multilingual neural machine translation with task-specific attention [C]//Proceedings of the International Conference on Computational Linguistics,2018: 3112-3122.

[28] Gu J, Wang Y, Chen Y, et al. Meta-learning for low-resource neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing,2018: 3622-3631.

[29] Wang Y, Zhang J, Zhai F, et al. Three strategies to improve one-to-many multilingual translation [C]//Proceedings of the Empirical Methods in Natural Language Processing,2018: 2955-2960.

[30] Kiperwasser E, Ballesteros M. Scheduled multi-task learning: From syntax to translation[J]. Transactions of the Association for Computational Linguistics, 2018: 225-240.

[31] Zaremoodi P, Haffari G. Neural machine translation for bilingually scarce scenarios: A deep multi-task learning approach[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics,2018: 1356-1365.

[32] Niu X, Rao S, Carpuat M. Multi-task neural models for translating between styles within and across languages[C]//Proceedings of the International Conference on Computational Linguistics,2018: 1008-1021.

[33] Su J, Zeng J, Xiong D, et al. A hierarchy-to-sequence attentional neural machine translation model [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2018, 26(3): 623-632.

[34] Zhang W, Hu J, Feng Y, et al. Refining source representations with relation networks for neural machine translation[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 1292-1303.

[35] Xiong H, He Z, Hu X, et al. Multi-channel encoder for neural machine translation[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 4962-4969.

[36] Shen Y, Tan X, He D, et al. Dense information flow for neural machine translation[C]//Proceedings of the North American Chapter of The Association for Computational Linguistics, 2018: 1294-1303.

[37] Werlen L M, Pappas N, Ram D, et al. Self-attentive residual decoder for neural machine translation[C]//Proceedings of the North American Chapter of The Association for Computational Linguistics, 2018: 1366-1379.

[38] Wang M, Xie J, Tan Z, et al. Neural machine translation with decoding history enhanced attention[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 1464-1473.

[39] Zhang B, Xiong D, Su J. Neural machine translation with deep attention[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 11.

[40] Luong T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [C]//Proceedings of the Empirical Methods in Natural Language Processing, 2015: 1412-1421.

[41] Chen K, Wang R, Utiyama M, et al. Syntax-directed attention for neural machine translation[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 4792-4799.

[42] Meng F, Tu Z, Cheng Y, et al. Neural machine translation with key-value memory-augmented attention [C]//Proceedings of the International Joint Conference on Artificial Intelligence, 2018: 2574-2580.

[43] Li X, Liu L, Tu Z, et al. Target foresight based attention for neural machine translation[C]//Proceedings of the North American Chapter of The Association for Computational Linguistics, 2018: 1380-1390.

[44] Huang P, Wang C, Huang S, et al. Towards neural phrase-based machine translation[C]//Proceedings of the International Conference on Learning Representations, 2018: 1-14.

[45] Wang C, Wang Y, Huang P, et al. Sequence modeling via segmentations[C]//Proceedings of the International Conference on Machine Learning, 2017: 3674-3683.

[46] Shi X, Knight K. Speeding up neural machine translation decoding by shrinking run-time vocabulary. [C]//Proceedings of Meeting of the Association for Computational Linguistics, 2017: 574-579.

[47] Zhang W, Huang L, Feng Y, et al. Speeding up neural machine translation decoding by cube pruning[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 4284-4294.

[48] Post M, Vilar D. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation[C]//Proceedings of the North American Chapter of The Association for Computational Linguistics, 2018: 1314-1324.

[49] Kuang S, Li J, Branco A, et al. Attention focusing for neural machine translation by bridging source and target embeddings[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2018: 1767-1776.

[50] Nguyen T Q, Chiang D. Improving lexical choice in neural machine translation[C]//Proceedings of the North American Chapter of The Association for Computational Linguistics, 2018: 334-343.

[51] Liu F, Lu H, Neubig G. Handling homographs in neural machine translation[C]//Proceedings of the North American Chapter of The Association for Computational Linguistics, 2018: 1336-1345.

[52] Wang L, Tu Z, Shi S, et al. Translating pro-drop languages with reconstruction models[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 4937-4945.

[53] Li Y, Li J, Zhang M. Adaptive weighting for neural machine translation[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 3038-3048.

[54] Zhang B, Xiong D, Su J, et al. Simplifying neural machine translation with addition-subtraction twin-gated recurrent networks[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 4273-4283.

[55] See A, Luong M, Manning C D, et al. Compression of neural machine translation models via pruning[C]//Proceedings of the Conference on Computational Natural Language Learning, 2016: 291-301.

[56] Shu R, Nakayama H. Compressing word embeddings via deep compositional code learning[C]//Proceedings of the International Conference on Learning Representations, 2018: 1-13.

[57] Xia Y, Tian F, Wu L, et al. Deliberation networks: Sequence generation beyond one-pass decoding[C]//Proceedings of the Neural Information Processing Systems, 2017: 1784-1794.

[58] Zhang X, Su J, Qin Y, et al. Asynchronous bidirectional decoding for neural machine translation[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 5698-5705.

[59] Geng X, Feng X, Qin B, et al. Adaptive multi-pass decoder for neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 523-532.

[60] Su J, Wu S, Xiong D, et al. Variational recurrent neural machine translation[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 5488-5495.

[61] Schulz P, Aziz W, Cohn T. A stochastic decoder for neural tachine translation[C]//Proceedings of Meeting of The Association for Computational Linguistics, 2018: 1243-1252.

[62] Lin J, Sun X, Ren X, et al. Deconvolution-based global decoding for neural machine translation[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 3260-3271.

[63] Zheng Z, Zhou H, Huang S, et al. Modeling past and future for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2018: 145-157.

[64] Kuang S, Xiong D. Fusing recency into neural machine translation with an inter-sentence gate model[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 607-617.

[65] Tu Z, Liu Y, Shi S, et al. Learning to remember translation history with a continuous cache[J]. Transactions of the Association for Computational Linguistics, 2018: 407-420.

[66] Maruf S, Haffari G. Document context neural machine translation with memory networks[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2018: 1275-1284.

[67] Kuang S, Xiong D, Luo W, et al. Modeling coherence for neural machine translation with dynamic and topic Caches[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 596-606.

[68] Wang X, Tu Z, Zhang M, et al. Incorporating statistical machine translation word knowledge into neural machine translation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2255-2266.

[69] Zhao Y, Wang Y, Zhang J, et al. Phrase table as recommendation memory for neural machine translation[C]//Proceedings of the International Joint Conference on Artificial Intelligence, 2018: 4609-4615.

[70] Zhao Y, Zhang J, He Z, et al. Addressing troublesome words in neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 391-400.

[71] Chen H, Huang S, Chiang D, et al. Improved neural machine translation with a syntax-aware encoder and decoder[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2017: 1936-1945.

[72] Li J, Xiong D, Tu Z, et al. Modeling source syntax for neural machine translation. [C]//Proceedings of Meeting of the Association for Computational Linguistics, 2017: 688-697.

[73] Wu S, Zhang D, Yang N, et al. Sequence-to-dependency neural machine translation[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2017: 698-707.

[74] Wu S, Zhang D, Zhang Z, et al. Dependency-to-dependency neural machine translation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 2132-2141.

[75] Ma C, Tamura A, Utiyama M, et al. Forest-based neural machine translation[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2018: 1253-1263.

[76] Zaremoodi P, Haffari G. Incorporating syntactic uncertainty in neural machine translation with a forest-to-sequence model[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 1421-1429.

[77] Gu J, Shavarani H, Sarkar A. Top-down tree structured decoding with syntactic connections for neural machine translation and parsing[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 401-413.

[78] Bastings J, Titov I, Aziz W, et al. Graph convolutional encoders for syntax-aware neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2017: 1957-1967.

[79] Li Q, Wong D, Chao L, et al. Linguistic knowledge-aware neural machine translation[J]. IEEE Transac-

tions on Audio, Speech, and Language Processing, 2018,26(12): 2341-2354.

[80] Gu J, Wang Y, Cho K, et al. Search engine guided neural machine translation[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 5133-5140.

[81] Zhang J, Utiyama M, Sumita E, et al. Guiding neural machine translation with retrieved translation pieces [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics, 2018: 1325-1335.

[82] Ugawa A, Tamura A, Ninomiya T, et al. Neural machine translation incorporating named entity[C]// Proceedings of the International Conference on Computational Linguistics, 2018: 3240-3250.

[83] Vaswani A, Shazzer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the Neural Information Processing Systems, 2017: 5998-6008.

[84] Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation[C]//Proceedings of the International Conference on Learning Representations, 2018: 1-13.

[85] Wang C, Zhang J, Chen H, Semi-autoregressive neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 479-488.

[86] Shaw P, Uszkoreit J, Vaswani A, Self-attention with relative position representations[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics, 2018: 464-468.

[87] Yang B, Tu Z, Wong D, et al. Modeling localness for self-attention networks[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 4449-4458.

[88] Shen T, Zhou T, Long G, et al. DiSAN: Directional self-attention network for RNN/CNN-Free language understanding[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 5446-5455.

[89] Zhang B, Xiong D, Su J, Accelerating neural transformer via an average attention network[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2018: 1789-1798.

[90] Li J, Tu Z, Yang B, et al. Multi-head attention with disagreement regularization[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 2897-2903.

[91] Shen T, Zhou T, Long G, et al. Bi-Directional Block Self-Attention for Fast and Memory-Efficient Sequence Modeling[C]//Proceedings of the International Conference on Learning Representations, 2018: 1-18.

[92] Dou Z, Tu Z, Wang X, et al. Exploiting deep representations for neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 4253-4262.

[93] Wang Q, Li F, Xiao T, et al. Multi-layer representation fusion for neural machine translation[C]//Proceedings of the International Conference on Computational Linguistics. 2018: 3015-3026.

[94] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the International Conference on Machine Learning, 2017: 1243-1252.

[95] Gehring J, Auli M, Grangier D, et al. A Convolutional encoder model for neural machine translation[C]// Proceedings of Meeting of the Association for Computational Linguistics, 2017: 123-135.

[96] Kaiser L, Gomez A N, Chollet F. Depthwise separable convolutions for neural machine translation[C]// Proceedings of the International Conference on Learning Representations, 2018: 1-10.

[97] Chen M X, Firat O, Bapna A, et al. The best of both worlds: Combining recent advances in neural machine translation[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2018: 76-86.

[98] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. Computer Vision and Pattern Recognition, 2016: 2818-2826.

[99] Domhan T. How much attention do you need A granular analysis of neural machine translation architectures[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, 1: 1799-1808.

[100] Wang M X. Neural machine translation with decoding history enhanced attention[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 1464-1473.

[101] Lakew S M, Cettolo M, Federico M. A comparison of transformer and recurrent neural networks on multilingual neural machine translation[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 641-652.

[102] Tang G, Muller M, Gonzales A R, et al. Why self-attention? A targeted evaluation of neural machine translation architectures[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 4263-4272.

[103] Ott M A, Auli M, Grangier D, et al. Analyzing uncertainty in neural machine translation[C]//Proceedings of the International Conference on Machine Learning, 2018: 3953-3962.

[104] Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation[C]//Proceedings of the International Conference on Learning Representations, 2018: 1-13.

[105] Tan X, Wu L, He D, et al. Beyond error propagation in neural machine translation: Characteristics of language also matter[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 3602-3611.

[106] Bahdanau D, Brakel P, Xu K, et al. An actor-critic algorithm for sequence prediction[C]//Proceedings of the International Conference on Learning Representations, 2017: 1-17.

[107] He D, Lu H, Xia Y, et al. Decoding with value networks for neural machine translation[C]//Proceedings of the Neural Information Processing Systems, 2017: 178-187.

[108] Wu L, Tian F, Qin T, et al. A study of reinforcement learning for neural machine translation[C]//Proceedings of the Empirical Methods in Natural Language Processing, 2018: 3612-3621.

[109] Ebrabimi J, Lowd D, Dou D. On adversarial examples for character-level neural machine translation[C]//Proceedings of the International Conference on Computational Linguistics, 2018: 653-663.

[110] Cheng Y, Tu Z, Meng F, et al. Towards robust neural machine translation[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2018: 1756-1766.

[111] Gu J, Im D J, Li V O. Neural machine translation with gumbel-greedy decoding[C]//Proceedings of the National Conference on Artificial Intelligence, 2018: 5125-5132.

[112] Yang Z, Chen W, Wang F, et al. Improving neural machine translation with conditional sequence generative adversarial nets[C]//Proceedings of the North American Chapter of Association for Computational Linguistics, 2018: 1346-1355.

**林倩**(1994—),**硕士研究生,主要研究领域为机器翻译、自然语言处理。**
E-mail: linqian17@stu.xmu.edu.cn

**苏劲松**(1982—),**通信作者,博士,副教授,主要研究领域为机器翻译、自然语言处理。**
E-mail: jssu@xmu.edu.cn

**刘庆**(1997—),**硕士研究生,主要研究领域为机器翻译、自然语言处理。**
E-mail: qingliu@stu.xmu.edu.cn