神经机器翻译综述

李亚超[1),2)] 熊德意[1)] 张民[1)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](西北民族大学甘肃省民族语言智能处理重点实验室 兰州 730030)

摘 要 机器翻译研究将源语言所表达的语义自动转换为目标语言的相同语义，是人工智能和自然语言处理的重要研究内容。近年来，基于序列到序列模型（Sequence-to-Sequence Model）形成一种新的机器翻译方法：神经机器翻译（Neural Machine Translation, NMT），它完全采用神经网络完成源语言到目标语言的翻译过程，成为一种极具潜力全新的机器翻译模型。神经机器翻译经过最近几年的发展，取得了丰富的研究成果，在多数语言对上逐渐超过了统计机器翻译方法。该文首先介绍了经典神经机器翻译模型及存在的问题与挑战；然后简单概括神经机器翻译中常用的神经网络；之后按照经典神经机器翻译模型、基础共性问题、新模型、新架构等分类体系详细介绍了相关研究进展；接着简单介绍基于神经网络的机器翻译评测方法；最后展望未来研究方向和发展趋势，并对该文做出总结。

关键词 机器翻译；神经机器翻译；注意力机制；循环神经网络；序列到序列模型；机器翻译评测

中图法分类号 TP18

A Survey of Neural Machine Translation

LI Ya-Chao[1)2)] XIONG De-Yi[1)] ZHANG Min[1)]

[1)](School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006)

[2)](Key Laboratory of National Language Intelligent Processing, Northwest Minzu University, Lanzhou 730030)

Abstract Machine translation is a subfield of artificial intelligence and natural language processing that investigates transforming the source language into the target language. Neural machine translation is a recently proposed framework for machine translation based purely on sequence-to-sequence models, in which a large neural network is used to transform the source language sequence into the target language sequence, leading to a novel paradigm for machine translation. After years of development, NMT has gained rich results and gradually surpassed the statistical machine translation (SMT) method over various language pairs, becoming a new machine translation model with great potential. In this paper, we systematically describe the vanilla NMT model and the different types of NMT models according to the principles of classical NMT model, the common and shared problems of NMT model, the novel models and new architectures, and other classification systems. First, we introduce the Encoder-Decoder based NMT as well as the problems and challenges in the model. In the vanilla NMT model, the encoder, implemented by a recurrent neural network (RNN), reads an input sequence to produce a fixed-length vector, from which the decoder generates a sequence of target language words. The biggest issue in the vanilla NMT model is that a sentence of any length needs to be compressed into a fixed-length vector that may be losing important information of a sentence, which is a bottleneck in NMT. Next,

本课题得到国家自然科学基金(61525205, 61432013, 61403269)、西北民族大学中央高校基本科研业务费专项资金资助项目(31920170154, 31920170153)、甘肃省高等学校科研项目资助(2016B-007).李亚超，男，1986年生，博士生，讲师，计算机学会(CCF)会员（73612M）主要研究领域为机器翻译和自然语言处理.E-mail: harry_lyc@foxmail.com.熊德意，男，1979年生，博士，教授，计算机学会会员（57174M），主要研究领域为自然语言处理、机器翻译、多语言信息获取.E-mail: dyxiong@suda.edu.cn.张民（**通讯作者**），男，1970年生，博士，教授，博士生导师，计算机学会会员（38729M），主要研究领域为机器翻译和自然语言处理.E-mail: minzhang@suda.edu.cn.

we summarize the neural networks used in NMT, including RNNs, convolutional neural networks (CNN), long short-term memory (LSTM) neural networks, gated recurrent neural networks, neural Turing machines (NTM), and memory networks, et al. Then, this paper introduces the current research situation of NMT in detail, including the attention-based NMT through attention mechanism, which is designed to predict the soft alignment between the source language and the target language, thus has greatly improved the performance of NMT; the character-level NMT model, aiming to solve the problems in the word-level NMT model, including character-level translation, subword-level translation, et al; the multilingual NMT, which has the ability to use a single NMT model to translate between multiple languages, including the one-to-many model, the many-to-one model and the many-to-many model; the problem of restriction in NMT, focusing on solving the very large target vocabulary in NMT, including the out-of-vocabulary (OOV) problems and how to address the long sentence problems in NMT; leveraging prior knowledge in NMT, for example, incorporating and effective utilization of the word reordering knowledge, the morphological features, the bilingual-dictionary, the syntactic information and the monolingual data into NMT; the low-resource NMT, which is a solution to the poor-resource training data conditions for some language pairs; the new paradigm for the NMT architectures, for example the multi-model NMT, the NMT model via non recurrent neural networks, and the advanced learning paradigm for NMT, such as generative adversarial networks (GAN) and reinforcement learning. Last, we summarize some successful evaluation methods of machine translation based purely on neural networks. Finally, the paper gives a future outlook on the development trend of NMT and summarize the key challenges and possible solutions.

Key words machine translation; neural machine translation; attention mechanism; recurrent neural network; sequence-to-sequence model; machine translation evaluation

1 引言

机器翻译研究如何利用计算机自动地实现不同语言之间的相互转换，是自然语言处理和人工智能重要研究领域，也是目前互联网常用服务之一。如 Google 翻译、百度翻译、微软 Bing 翻译等，都提供了多种语言之间的在线翻译服务。虽然机器翻译译文质量与专业译员相比仍有较大差距，但是在一些对译文质量要求不太高的场景下，或者是在特定领域翻译任务上，机器翻译在翻译速度上具有明显优势，仍然得到广泛应用。鉴于机器翻译的复杂性和应用前景，学术界和产业界都把该领域作为重点研究方向，成为当前自然语言处理最活跃的研究领域之一。

1957 年，Rosenblatt 提出了感知机（Perceptron）算法[1]，这是一种最简单的神经网络。早期的感知机，因其结构简单，不能处理线性不可分问题，造成了该研究长期的低潮期。20 世纪 80 年代以后，反向传播算法（Backpropagation, BP）被引入到多层感知机（Multilayer Perceptron, MLP），也叫前馈神经网络（Feedforward Neural Network, FNN）。此后，在 Hinton、LeCun、Bengio 等人推动下，神经网络重新引起人们关注。2006 年，Hinton 等人[2]通过逐层预训练方法解决了神经网络训练难题，随后由于计算能力提高，如并行计算、图形处理器（Graphics Processing Unit, GPU）的广泛应用，神经网络在学术界和产业界都得到高度重视。近年来，神经网络在图像识别[3]、语音识别[4]等领域取得巨大成功，同时学者们也将该技术应用在自然语言处理任务上，如语言模型、词语表示、序列标注等任务[5]，并取得了令人鼓舞的成绩。

机器翻译相关研究，在多种语言对上，神经机器翻译已经逐渐超过短语统计机器翻译。Junczys-Dowmunt 等人[6]采用联合国语料库（United Nations Parallel Corpus v1.0），在 30 个语言对上对神经机器翻译和短语统计机器翻译进行对比，神经机器翻译在 27 个语言对上超过了短语统计机器翻译方法。与汉语相关的，如中英、中俄、中法之间翻译任务上，神经机器翻译高出 6-9 个 BLEU 值（Bilingual Evaluation Understudy, BLEU）。另外，在 2016 年机器翻译研讨会（Workshop on Machine Translation, WMT）上，爱丁堡大学开发的神经机器翻译系统在英语到德语翻译任务上，超过基于短语、基于句法的统计机器翻译[7]。在大规模计算能力支持下，百度公司采用深层次神经网络架构，在

WMT 2014 英语到法语翻译任务上，首次超过统计机器翻译方法，取得了最好的成绩[8]。在产业界，Google 翻译在部分语言上已采用神经机器翻译代替统计机器翻译对外提供服务[9]。著名的商用机器翻译公司 Systran 同样开发出相应的神经机器翻译系统，涵盖了 12 种语言 32 个语言对[10]。在国内，搜狗公司、小牛翻译也在积极开发神经机器翻译系统。目前，神经机器翻译不仅在学术界得到广泛关注，产业界也积极地探索该方法的商用价值。

由于自然语言的多样性和复杂性，将一种语言恰如其分地翻译为另外一种语言仍然困难重重。目前，在大规模语料和计算能力条件下，神经机器翻译展现出巨大潜力，已经发展成为一种新的机器翻译方法。这种方法仅需要双语平行语料，便于训练大规模翻译模型，不仅具有很高的研究价值，同时也具有很强的产业化能力，成为当前机器翻译研究的前沿热点。

本文剩余部分结构如下：第 2 部分讲述经典神经机器翻译模型及其面临的问题和挑战；第 3 部分概述在神经机器翻译中常用的神经网络及其特点；第 4 部分详细介绍神经机器翻译研究进展；第 5 部分为基于神经网络的机器翻译评测方法概述；第 6 部分展望神经机器翻译未来研究方向；第 7 部分为本文小结。

2 经典神经机器翻译模型及其问题与挑战

2.1 经典神经机器翻译模型

统计机器翻译把翻译问题等同于求解概率问题，即给定源语言 s，求目标语言 t 的条件概率 p。选取好翻译模型后，从双语平行语料中学习到这些模型的参数。当输入源语言时，通过学习到的模型最大化上述条件概率来获得最优翻译结果。

神经机器翻译依据上述基本思想，在翻译建模上则完全采用神经网络实现了源语言到目标语言的直接翻译。这种翻译思想提出的很早，在 20 世纪 90 年代，有学者采用小规模语料实现了基于神经网络的翻译方法[11-12]，由于语料资源和计算能力限制，并没有得到相应关注。在深度学习热潮兴起之后，神经网络常用于统计机器翻译的语言模型、词语对齐、翻译规则抽取等[13]。直到 2013 年，基于神经网络的翻译方法被 Kalchbrenner 和 Blunsom[14]重新提出，展现出了巨大的应用潜力。随后，Sutskever[15]、Cho[16-17]、Jean[18-19]等人分别实现相应的完全基于神经网络的机器翻译模型。这些属于经典神经机器翻译模型，本质上是序列到序列模型，不仅可以用于机器翻译，还可以应用到问答系统、文本摘要等其他自然语言处理任务。

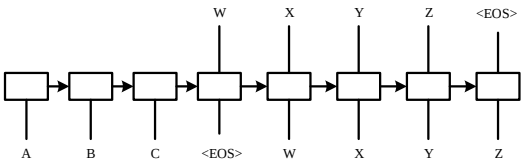

图 1 端到端模型

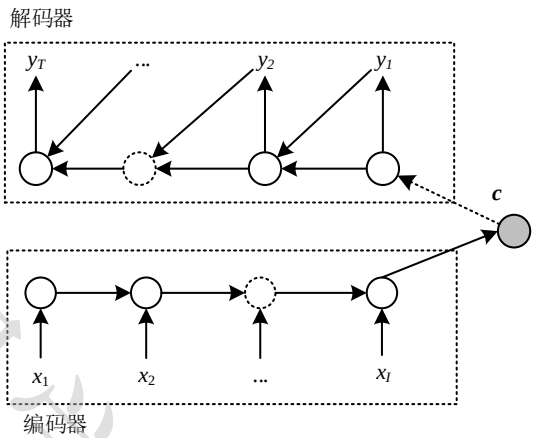

图 2 编码器解码器模型

与统计机器翻译的离散表示方法不同，神经机器翻译采用连续空间表示方法（Continuous Space Representation）表示词语、短语和句子。在翻译建模上，不需要词对齐、翻译规则抽取等统计机器翻译的必要步骤，完全采用神经网络完成从源语言到目标语言的映射。这种翻译模型大致可以分为两种，第一种是 Google 提出的翻译模型[15]，另外一种是蒙特利尔大学提出的翻译模型[16]，两种模型在原理上非常相近。第一种如图 1 所示，模型输入“A”、“B”、“C”，在输入条件下依次生成输出“W”、“X”、“Y”、“Z”，“<EOS>”为人为加入的句子结束标志。在翻译中，输入为源语言，输出为目标语言，称为端到端模型（End-to-End Model）[15]。

另外一种称为编码器解码器模型（Encoder-Decoder Model）[16]，在下文中对这个模型作详细介绍。其中编码器读取源语言句子，将其编码为维数固定的向量；解码器读取该向量，依次生成目标语言词语序列，如图 2 所示。

编码器解码器模型由三部分组成，输入 $\boldsymbol{x}$，隐藏状态 $\boldsymbol{h}$，输出 $\boldsymbol{y}$。编码器读取输入 $\boldsymbol{x} = (x_1, x_2, \ldots, x_I)$，将其编码为隐藏状态 $\boldsymbol{h} = (h_1, h_2, \ldots, h_I)$，当采用循环神经网络（RNN）时：

$$h_i = f(x_i, \quad (1)$$

$$\boldsymbol{c} = q(\{h_1, \ldots \quad (2)$$

$\boldsymbol{c}$ 是源语言句子表示，f 和 q 是非线性函数。

解码器在给定源语言表示 $\boldsymbol{c}$ 和前驱输出序列 $\{y_1, \ldots, y_{t-1}\}$，生成目标语言词语 y_t，定义如下：

$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t | \{y_1, \ldots, y_{t-1}\}, \boldsymbol{c}) \quad (3)$$

$\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$，当采用循环神经网络时：

$$p(y_t | \{y_1, \ldots, y_{t-1}, \boldsymbol{c}\}) = g(y_{t-1}, s_t, \boldsymbol{c}) \quad (4)$$

g 是非线性函数用来计算 y_t 的概率，s_t 是循环神经网络的隐藏状态，$s_t = f(s_{t-1}, y_{t-1}, \boldsymbol{c})$。

编码器和解码器可以进行联合训练，形式如下：

$$\mathcal{L}(\theta) = \max_{\theta} \frac{1}{N} \sum_{n=1}^{N} log\, p_\theta(y_n | x_n) \quad (5)$$

θ 是模型的参数，通过梯度下降法计算，(x_n, y_n) 是双语句对。

编码器解码器模型是通用的框架，可以由不同的神经网络实现，如长短时记忆神经网络（LSTM）[15]、门控循环神经网络[16]（Gated Recurrent Neural Networks）等。

神经机器翻译仅需要句子级平行语料，单纯采用神经网络实现翻译过程，便于训练大规模的翻译模型，具有很高实用价值。经验证，其翻译效果接近或达到基于短语的统计机器翻译方法[6]。在一些译文细粒度评价指标上神经机器翻译也具有很大优势，比如，Bentivogli 等人[20]在 2015 年口语翻译国际研讨会（The International Workshop on Spoken Language Translation, IWSLT）英语到德语翻译评测任务上，对短语统计机器翻译和神经机器翻译的译文进行了详细的对比分析。神经机器翻译译文中形态错误减少了 19%，词汇错误减少了 17%，词语调序错误减少了 50%。词语调序错误中，动词调序错误减少了 70%。

基于以上分析，神经机器翻译在多个评价指标上逐渐超过了统计机器翻译方法，成为一种非常具有潜力的机器翻译模型。

2.2 神经机器翻译与统计机器翻译异同

机器翻译方法可以分为基于规则的机器翻译，基于实例的机器翻译，以及统计机器翻译[21]。从 20 世纪 90 年代以来，随着语料库规模扩大，以及计算能力提高，统计机器翻译成为这个时期的主流方法。本文只论述神经机器翻译与统计机器翻译的相同点与不同点。

把机器翻译看作求解概率问题，是统计机器翻译的核心思想。在这基本思想上，统计机器翻译和神经机器翻译是一致的，不同之处在于具体实现方式上。

统计机器翻译根据贝叶斯原理对 p 进行扩展得到以下公式：

$$p(\boldsymbol{t}|\boldsymbol{s}) = \frac{p(\boldsymbol{t})p(\boldsymbol{s}|\boldsymbol{t})}{p(\boldsymbol{s})} \quad (6)$$

公式的分母表示源语言句子概率，在具体任务上是固定值。因此求 $p(\boldsymbol{t}|\boldsymbol{s})$ 的最大值，等同于寻找 $\hat{t}$，使公式右边的乘积最大，即：

$$\hat{t} = arg\max_{t} p(\boldsymbol{t})p(\boldsymbol{s}|\boldsymbol{t}) \quad (7)$$

其中 $p(\boldsymbol{t})$ 是语言模型，$p(\boldsymbol{s}|\boldsymbol{t})$ 是翻译模型。在统计机器翻译中可以进一步分解为多个子模块，如语言模型、翻译模型、调序模型等，并通过对数线性模型结合在一起，共同完成翻译过程。

神经机器翻译则采用神经网络实现源语言到目标语言的直接翻译。从整体上看，该方法类似一个黑箱结构，对于统计机器翻译的必要部分，如词对齐、语言模型、翻译模型等都是具备的，采用一种隐含的方式实现。两者不同之处如下所示：

（1）词对齐建模：词对齐对源语言和目标语言词语之间的对应关系建模，是统计机器翻译的重要部分。经典神经机器翻译模型并不需要词对齐步

骤，基于注意力机制（Attention Mechanism）的神经机器翻译[22]，在解码时能够动态地获得与生成词语相关的源语言词语信息。虽然通过注意力机制可以得到词对齐信息，但是这种词对齐与统计机器翻译词对齐相比，包含的信息较少，对齐效果也较弱。

（2）翻译效果对比：神经机器翻译在生成译文时利用了源语言信息和已生成译文信息，等同于将多个模块无缝的融合在一起。实验证明，神经机器翻译译文流利度要优于统计机器翻译，对于统计机器翻译难以有效处理的复杂结构调序和长距离调序问题，也能够较好地处理[20]。但是在翻译忠实度上，神经机器翻译要差一些[23]。

除以上所述，神经机器翻译与统计机器翻译的不同之处如表 1 所示。NMT、SMT 分别表示神经机器翻译和统计机器翻译。

表 1 NMT 与 SMT 差异

评价指标	NMT	SMT
表示方法	连续	离散
模型	非线性	对数线性
模型大小	小	大
训练时间	长	短
模型可解释性	弱	强
内存占用	小	大
GPU	必须	非必须
增量式训练	支持	不支持

2.3 问题与挑战

基于编码器解码器结构的神经机器翻译是一种通用的模型，并不完全针对机器翻译任务本身而设计，导致神经机器翻译仍然存在一些问题亟待解决。

（1）受限制的词典大小和句子长度：神经机器翻译要求双语词典大小固定，考虑到训练复杂度，通常将词典大小、句子长度限制在较小范围[19]。致使神经机器翻译面临更加严峻的未登录词、长句子翻译问题。因此，实现词典大小无限制，或者是能够高效地处理未登录词问题，同时对较长句子也能够有效翻译，是神经机器翻译需要解决的基本问题。

（2）难以高效利用外部先验知识：神经机器翻译只采用双语训练数据，不要求额外先验知识，如大规模单语语料、标注语料、双语词典等。另外，神经机器翻译的结构特点决定了采用外部资源是很困难的。单语语料、标注语料、双语词典等资源在统计机器翻译中可以显著提高翻译质量[24]，而这些先验知识在神经机器翻译中并没有得到充分应用。因此，高效利用外部先验知识具有很高实用价值，成为亟待解决的问题。

（3）注意力机制有待进一步完善：注意力机制是对神经机器翻译的重大改进[22]，不足之处是生成目标语言词语时，并没有考虑到历史注意力信息，且约束机制较弱。此外，在一些情况下，生成目标语言词语时并不需要过多关注源语言信息，比如汉英翻译中，要生成虚词“The”时，应该更多关注目标语言相关信息。除以上所述，神经机器翻译中存在过度翻译（Over Translation）和翻译不充分（Under Translation）问题[23]，同样需要完善现有注意力机制。在神经机器翻译中，完善注意力机制是研究的热点和难点。

（4）神经网络架构单一：基于编码器解码器的神经机器翻译在架构上较为简单，仅能捕捉句子级词语信息。目前有学者通过在神经机器翻译中融合重构（Reconstruction）思想，提高翻译忠实度[25]；采用半监督学习方法，有效利用源语言和目标语言单语语料[26]；采用变分神经机器翻译（Variational Neural Machine Translation, VNMT），替代神经机器翻译[27]；通过添加外部记忆，提高神经机器翻译的建模能力[28]。综上所述，如何优化翻译模型架构是神经机器翻译所面临的重要挑战。

3 神经网络在机器翻译中的应用

神经网络依据拓扑结构特点可以分成多种类型，如前馈神经网络，卷积神经网络（Convolutional Neural Network），循环神经网络等。本文只介绍一些在机器翻译、句法分析、序列标注等自然语言处理任务上常用的神经网络，并对其在机器翻译上的应用作简要概述。

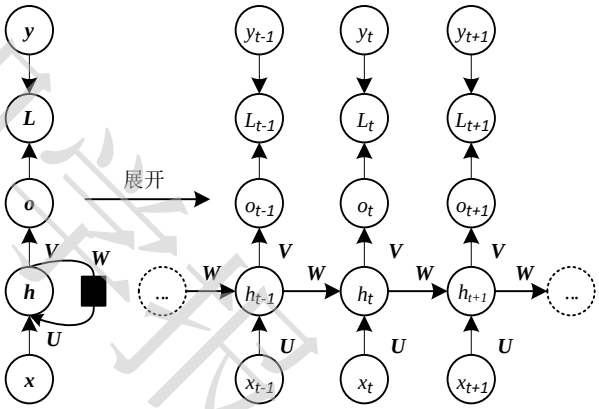

图 3 循环神经网络结构图

3.1 循环神经网络

循环神经网络主要用于处理序列数据，特别是对变长序列数据有着较好的处理能力[29]，神经机器翻译多数采用循环神经网络实现。如图 3 所示[30]。

$\boldsymbol{x}=\{x_1, x_2, \ldots, x_T\}$表示变长序列数据，在每个时间点 t 上，隐藏状态h_t由以下公式进行更新：

$$h_t = f(h_{t-1}, x_t) \tag{8}$$

f 是非线性函数。

通过循环神经网络将输入 $\boldsymbol{x}$ 映射到输出 $\boldsymbol{o}$。$\boldsymbol{y}$ 是模型所要达到的目标序列（通常由训练语料给出），L 是损失函数（Loss Function），U 为输入到

隐藏层的权重矩阵，W 为隐藏层到隐藏层的权重矩阵，V 是隐藏层到输出的权重矩阵，时间序列 t 范围为[1, T]，整个网络通过如下进行更新：

$$a_t = Wh_{t-1} + Ux \tag{9}$$

$$h_t = \tanh \tag{10}$$

$$o_t = Vh \tag{11}$$

$$\hat{y}_t = \mathrm{softmax}(o_t) \tag{12}$$

循环神经网络使得不同长度的输入序列，其输入向量维数都相同，并且在每个时间点上可以采用相同的变换函数和参数，更适合处理变长序列数据。另外，循环结构在理论上能够捕捉到所有前驱状态，这在一定程度上解决了长距离依赖问题。

3.2 循环神经网络的变形结构

将循环神经网络展开后可以采用反向传播算法训练，称为时间反向传播（Backpropagation Through Time, BPTT），在实际应用中会产生梯度消失问题（Vanishing Gradient Problem）[31]。长短时记忆神经网络[32]是循环神经网络的变形结构，采用了更加高效的遗忘和更新机制，具有与循环神经网络相似的结构和优点，且性能更好。

门限循环单元（Gated Recurrent Units, GRU）[16,33]将长短时记忆循环单元的输入门和遗忘门合并成更新门（Update Gate），又引入了重置门（Reset Gate），用更新门控制当前状态需要遗忘的历史信息和接受的新信息，用重置门控制候选状态中有多少信息是从历史信息中得到。该结构是对长短时记忆神经网络的简化，效果与后者相近，并降低了计算量。

递归神经网络（Recursive Neural Network，Recursive NN）是循环神经网络的变形结构，以树形结构进行组织，用于结构化预测和表示，适合表示自然语言句法结构[34]。

3.3 带记忆的神经网络

神经网络没有外部记忆（External Memory），对变量和数据长时间存储能力很弱，与外部信息交互很困难[35]。Graves 等人[36]将循环神经网络与外部记忆耦合，称为神经图灵机（NTM）。这种模型类似图灵机，并具有神经网络的优势，能够采用梯度下降法训练。除此之外，Weston 等人[37]提出了记忆网络（Memory Networks），包含一个长时记忆组件（Long-term Memory Component），能够读取和写入，在具体任务中可以作为知识库使用。

这些带外部记忆的神经网络能够方便地利用外部资源，增加了神经网络与外部资源交互能力，同时也提高了可解释性和记忆能力。

4 神经机器翻译研究进展

神经机器翻译源于序列到序列模型，已经发展成为一种全新的机器翻译方法。本节首先介绍基于注意力的神经机器翻译，这是对经典神经机器翻译模型的重大改进，然后对神经机器翻译关键技术研究进展进行分析、对比和总结。分类标准和分类体系如图 4 所示。

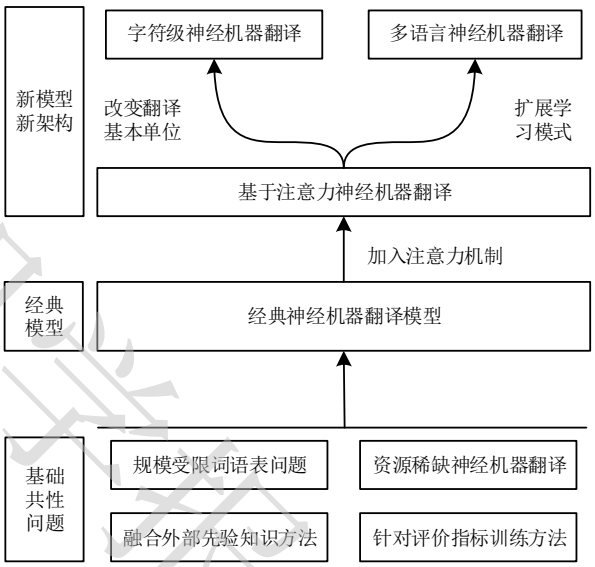

图 4 神经机器翻译模型分类体系

4.1 神经机器翻译注意力机制研究进展

注意力机制[22]是对经典神经机器翻译模型的完善，通过改进源语言表示方式，在解码中动态生成源语言相关信息，从而极大地提高了翻译效果，成为目前的主流方法，也是当前研究热点之一。

4.1.1 注意力机制及存在问题

基于注意力的神经机器翻译将源语言句子编码为向量序列，而不是一个固定向量，在生成目标语言词语时，能够利用与生成该词相关的源语言词语信息，所对应词语在源语言中可以连续存在，也可以离散分布[22]，如图 5 所示。注意力机制实现的双语词汇对应关系称为软对齐（Soft-alignment）。

与统计机器翻译硬对齐（Hard-alignment）方法相比，该方法对目标语言词语和源语言词语对齐长度不作限制，可以避免硬对齐方法中的对空问题。

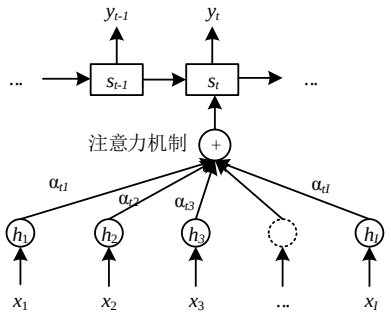

图 5 注意力机制图示

当采用注意力机制时，公式 4 重新定义为：

$$p(y_t|\{y_1,\dots,y_{t-1},\boldsymbol{x}\}) = g(y_{t-1},s_t,c_t) \quad (13)$$

s_t是 t 时刻的隐藏状态，$s_t = f(s_{t-1},y_{t-1},c_t)$。

上下文向量（Context Vector）c_t 依赖于源语言编码序列$(h_1, h_2,\dots, h_I)$，h_i 是第 i 个输入词的编码，计算方法如下：

$$c_t = \sum_{j=1}^{I} \alpha_{tj} h_j \quad (14)$$

是 h_j 的权重，计算方法如下：

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{I}\exp(e_{tk})} \quad (15)$$

$e_{tj} = a(s_{t-}$是对齐模型，表示 t 时刻的生成词语与第 j 个源语言词语的匹配程度。

基于注意力的神经机器翻译在解码时能够动态获取源语言相关信息，显著提升了翻译效果[22]，是神经机器翻译重要研究进展之一。

注意力机制是一种无监督的模型，不同时刻的注意力之间没有明显的约束条件，且求注意力分配权重时，需要计算源语言句子中所有词语的权重，很耗费计算资源。设计更加完善的注意力机制，成为当前研究热点，并取得了一系列重要成果。

4.1.2 减少注意力计算量方法

注意力机制存在计算量较大问题。为了减少计算量，Xu 等人[38]在图像描述生成任务上，将注意力分为软注意力（Soft Attention）和硬注意力（Hard Attention），前者指给原图像所有区域分配权重，计算量较大；后者指仅仅注意部分原图像区域，可以减少计算量。

根据上述思想，Luong 等人[39]提出了局部注意力（Local Attention）模型，是对全局注意力（Global Attention）的改进，能够减少计算量。全局注意力在计算上下文向量 c_t 时，要考虑源语言的所有编码序列，与 Bahdanau 等人[22]提出的注意力机制类似，同样比较耗费计算量。局部注意力仅需关注源语言编码中一个小的上下文窗口，可以显著减少计算量。该方法核心在于从源语言找到一个与生成词语相关的对齐位置，在计算上下文向量 c_t 时，以该对齐点为中心，选取大小固定的窗口计算。

局部注意力在生成上下文向量时只关注源语言小部分区域，把无关信息过滤掉，适合长句子翻译。在 WMT 2014 英语到德语翻译上，局部注意力相比全局注意力提高了 0.9 个 BLEU 值。在长句子翻译实验上，局部注意力方法随着句子长度增加，翻译质量并没有降低。另外，在亚琛工业大学（RWTH Aachen）英德词对齐语料上，局部注意力词对齐错误率为 34%，全局注意力词对齐错误率为 39%。

4.1.3 有监督注意力机制

有监督注意力机制为利用高质量的先验词对齐知识指导注意力机制。基于以下事实，注意力机制在预测目标语言词语对应的源语言词语时并没有利用该词语自身信息，是一种无监督的学习模型，词对齐质量较差。而这个问题在统计机器翻译词对齐中已经得到很好处理，词对齐质量很高。

Liu 等人[40]根据以上思想提出采用统计机器翻译词对齐信息作为先验知识指导注意力机制的方法。基本思想很简单：首先，利用 GIZA++[21]获取训练语料词对齐信息；然后，在模型训练中，统计机器翻译词对齐作为先验知识指导注意力机制，使得基于注意力的词对齐尽可能与统计机器翻译的词对齐一致；最后，在测试过程中不需要先验词对齐信息。

实验采用 2008 年美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）举办的汉英机器翻译评测语料，相比基于注意力神经机器翻译，该方法提高了 2.2 个 BLEU 值。在清华词对齐语料[41]上，GIZA++词对齐错误率为 30.6%，基于注意力神经机器翻译词对齐错误率为 50.6%，该方法词对齐错误率为 43.3%。可以看出有

监督机制可以显著提高注意力机制词对齐质量，但是与统计机器翻译词对齐相比仍有较大差距，注意力机制仍有改进空间。

4.1.4 融合统计机器翻译词对齐信息

注意力机制对源语言和目标语言词语对应关系建模，是无监督的模型，没有利用任何先验知识和约束机制[42]。统计机器翻译词对齐包含了丰富的信息，质量相对较高。如在 IBM 模型[43]中，位变模型（Distortion Model）用于控制词语的重排序，繁衍模型（Fertility Model）用于控制一个源语言词语可以对应目标语言词语的数量，而注意力机制缺少这些约束信息。根据以上所述，将统计机器翻译词对齐信息引入注意力机制是一种可行的方法，这方面的工作主要有以下几种。

Feng 等人[44]将位变模型、繁衍模型思想引入基于注意力的神经机器翻译，实验采用 NIST 汉英翻译语料，相比基线系统提高了 2.1 个 BLEU 值，同时也能够提高词对齐效果。该方法比较重要的贡献是借助统计机器翻译的繁衍模型，在一定程度上缓解了过度翻译问题。Cohn 等人[45]则在注意力机制中融合了更多的结构化偏置（Structural Biases）信息，包括位置偏置（Position Bias）、马尔科夫条件（Markov Condition）、繁衍模型、双语对称（Bilingual Symmetry）等信息。实验在罗马尼亚语、爱沙尼亚语、俄语、汉语到英语四个语言对上进行，其中汉英翻译采用 BTEC 语料库（Basic Travel Expression Corpus, BTEC），相比基于注意力神经机器翻译，提高了 3 个 BLEU 值，而其余实验效果并不显著。Zhang 等人[46]将位变模型显式地集成到注意力机制中，使得该机制同时获得源语言的词语信息和词语重排序（Word Reordering）信息。在较大规模的汉英语料上能够显著提高翻译质量和词对齐质量。

4.1.5 过度翻译和翻译不充分问题

过度翻译指一些词或短语被重复地翻译，翻译不充分指部分词或短语没有被完整地翻译。该问题在神经机器翻译中普遍存在，包括基于注意力的神经机器翻译。

上述问题部分原因在于神经机器翻译并没有很好的机制来记忆历史翻译信息，比如已翻译词语信息和未翻译词语信息，从公式 13-15 可以看出。在这方面研究中，Tu 等人[23]提出的覆盖（Coverage）机制是很重要的研究成果。该方法将统计机器翻译的覆盖机制引入基于注意力神经机器翻译。设计了一种覆盖向量，用于记录翻译过程的历史注意力信息，能够使注意力机制更多地关注未翻译词语，并降低已翻译词语权重。覆盖机制是统计机器翻译常用的方法，用于保证翻译的完整性。在神经机器翻译中，直接对覆盖机制建模是很困难的，Tu 等人通过在源语言编码状态中增加覆盖向量，显式地指导注意力机制的覆盖度。这种方法可以缓解过度翻译和翻译不充分问题，效果很明显。虽然没有完全解决该问题，但仍然是对注意力机制的重大改进。

该问题的另外一种解决方法是在翻译过程中控制源语言信息和目标语言信息对翻译结果的影响比例。这种思想很直观，在翻译过程中源语言上下文和目标语言上下文分别影响翻译忠实度和流利度。因此，当生成实词时应多关注源语言上下文，生成虚词时应更多依靠目标语言上下文。这就需要一种动态手段控制翻译过程中两种信息对翻译结果的影响，而这种控制手段是神经机器翻译所缺少的。这方面典型工作为 Tu 等人[47]提出的上下文门（Context Gate）方法，在保证翻译流利度同时，也确保了翻译的忠实度。覆盖机制和上下文门能够结合在一起，互为补充。覆盖机制能够生成更好的源语言上下文向量，着重考虑翻译充分性；上下文门则能够根据源语言、目标语言上下文的重要程度，动态控制两种信息对生成目标语言词语的影响比重。

过度翻译和翻译不充分问题是神经机器翻译存在的问题之一，在商用神经机器翻译系统中仍然存在该问题，需要更加深入研究。

4.1.6 融合外部记忆方法

在神经网络中增加外部记忆[35-36]，解码时与之交互，可以扩展神经网络的表达能力。因为外部记忆可以将当前时刻的重要中间信息存储起来，用于后续时刻，以此增强神经网络的长时记忆能力。在一些任务上可以达到并超过传统的循环神经网络和长短时记忆神经网络[37]。

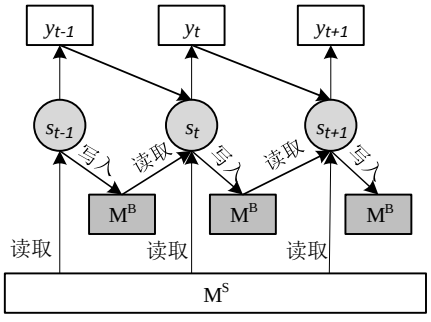

图 6 MEMDEC 解码方法图示

外部记忆应用在神经机器翻译的重要工作是 Wang 等人[28]提出的 MEMDEC 解码方法。该方法首先定义一个 n 大小的外部记忆，n 表示记忆单元个数，m 表示记忆单元大小，在解码过程中可以读取和写入信息，记忆单元的读取和写入类似神经图灵机的读写机制[36]。在解码中，将当前时刻的目标语言信息、源语言信息和解码器状态信息写入记忆里，并在下一时刻读取。如图 6 所示，M^S 为源语言记忆，即源语言表示，M^B 为外部记忆，s_t 表示隐藏状态，y_t 表示在 t 时刻生成的目标语言词语。

这种方法在记忆里选择性地存储可用于后续时刻的中间状态信息，在一定程度上弥补了注意机制的不足，能够更好地扩展神经机器翻译模型的表达能力及增强长距离依赖效果。

4.2 字符级神经机器翻译

字符级神经机器翻译（Character Level NMT）是为了解决未登录词、词语切分、词语形态变化等问题提出的一种神经机器翻译模型，主要特点是减小了输入和输出粒度。不同粒度词语切分示例如图 7 所示，空格表示词语之间切分，短线表示字符、亚词（Subword）之间切分。

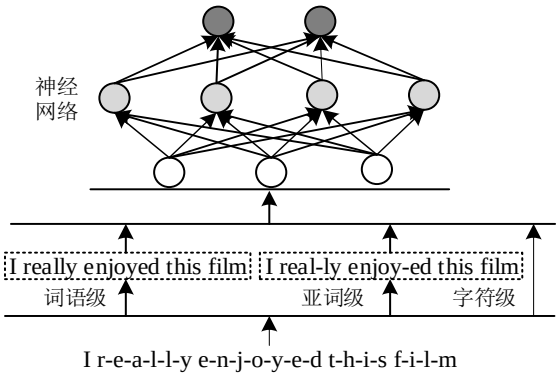

图 7 词语粒度示例

4.2.1 词语编码方案

多数神经机器翻译模型都以词语作为翻译基本单位，存在未登录词、数据稀疏，以及汉语、日语等语言中的分词问题。此外，在形态变化较多的语言中，如英语、法语等语言，以词为处理基本单位时，丢失了词语之间的形态变化、语义信息。如英语单词，“run”，“runs”，“ran”，“running”被认为是四个不同的词，忽略了他们有着共同的前缀“run”。为了解决上述问题，学者们提出了不同的词语编码方案，根据粒度划分可以归为以下两种：

（1）字符编码方案。对于英语、法语等拼音文字来说字符是组成词语的基本单位，在语言处理中能够以字符为单位建模。这方面工作很早就开始研究，比如字符级神经网络语言模型[48]。该方案同时也存在不足，比如编码粒度过小，适合英语、法语等字符数量相近的语言之间的翻译，如果用在英语到汉语翻译上会出现诸多问题。

（2）亚词编码方案。亚词编码方案选用的翻译基本单位介于字符和词语，可以得到两种方案的共同优势。词素的粒度同样介于字符和词语之间，不足之处是跟特定语言相关，限制了应用的通用性。因此，亚词通常采用 BPE 编码（Byte Pair Encoding, BPE）得到[49]，该方案将经常结合的字符组合看作是一个单位，比如词语“dreamworks interactive”，可以切分成“dre + am + wo + rks/ in + te + ra + cti + ve”序列，方法简单有效，适应性强。

4.2.2 半字符级神经机器翻译

半字符级神经机器翻译是编码器或者解码器的一端采用字符，另外一端采用亚词或者词语。这种方案是字符级和词语级编码的折中方案。

源语言端为亚词，目标语言端为字符或亚词，代表工作为 Chung 等人[50]提出的字符级解码方法。该方法中源语言翻译基本单位为亚词，通过 BPE 编码得到，目标语言以字符形式生成，编码器和解码器均采用循环神经网络实现。实验采用 WMT 2015 语料，源语言为英语，目标语言分别为捷克语、德语、法语、俄语等四个语言对。相比亚词级解码，字符级解码均取得了最好的翻译效果。在上述工作中发现了一些特点：（1）注意力机制能够实现字符到亚词、词语之间的对齐；（2）目标语言未登录词处理效果较好，因为字符级解码可以对任何词语建模；（3）解码中，字符序列显著长于亚词序列，但是在该实验中两者翻译效果相近。

源语言为字符，目标语言为词语是本节论述的另外一种形式。Costa-jussa 等人[51]将编码器的查找表（用于实现词语到词向量转换）替换为一个卷积神经网络，从而实现字符到词语的映射，解码器仍采用词语级解码。这种方法在源语言端采用字符级编码，能够捕捉到所有词语表达形式，消除了未登录词问题。这类工作同样应用在汉语、日语等需要分词的语言上。Su 等人[52]采用基于格循环神经网络（Lattice-based RNN），对汉语采用基于字的输入，通过词网格对词语的不同切分形式进行表示，并作

为编码器的输入，以此处理汉英翻译中汉语词语切分问题；解码时，仍然以词语形式生成英语翻译结果。与之类似，Yang 等人[53]则在编码器端采用行卷积（Row Convolution）神经网络，自动地从输入的字符序列学习到词语信息。这两种方法都以字符作为输入，可在源语言端减少未登录词问题，适合源语言需要分词的翻译任务，不足之处是目标语言端仍然为词语。

4.2.3 字符级神经机器翻译

字符级神经机器翻译要求输入和输出均以字符为基本单位。这类方法通过在编码器、解码器上增加字符到词语之间映射机制，从而实现字符序列输入和输出。

Ling 等人[54]在编码器上增加字符到词语映射，实现字符级输入，解码时生成目标语言字符序列，注意力机制关注源语言词语序列。这种方法人为在双语语料中加入“SOS”、“EOS”，分别表示句子开始、结尾标志；加入“SOW”、“EOW”，分别表示词语开始和结尾标志。用循环神经网络实现字符到词语的映射，构建基于字符的词语表示。在解码中实现字符级输出，包括词语和句子的开始、结尾标志。当产生“EOS”表示生成一个完整的句子，产生“EOW”表示生成一个完整的词语。通过这种方法实现了字符级的输入和输出。

Lee 等人[55]将字符向量（Character Embeddings）序列输入卷积神经网络，其输出分成长度固定的切分序列，对每个切分应用最大池化（Max-pooling）操作，得到切分编码（Segment Embeddings）。该切分编码作为自动学习的语义单元，输入编码器。在解码器中，注意力机制关注源语言切分编码序列，并生成目标语言字符序列。

这两种方法主要不同之处在于源语言语义基本单位，第二种方法为自动学习的语义单元，长度固定，而第一种方法中语义单元为词语。这类方法主要特点是在源语言端用神经网络实现字符到词语的映射，从而实现字符级输入，并消除了未登录词问题；在目标语言端，根据切分标志判断词语和句子边界。

4.3 多语言神经机器翻译

多语言机器翻译，区别于通常一种语言到另外一种语言的一对一翻译，能够采用一个模型完成多种语言之间翻译。基于神经网络的多语言机器翻译源于序列到序列学习和多任务学习（Multi-task Learning）[56-57]，从类型上可以分为单语到多语翻译、多语到单语翻译，以及多语到多语翻译。

4.3.1 单语到多语翻译

单语到多语翻译是源语言只有一个，而目标语言有多个的机器翻译方法。Dong 等人[57]首次将多任务学习引入序列到序列学习，实现了一种单语到多语的神经机器翻译方法。该方法在编码器解码器上增加了多任务学习模型，源语言采用一个编码器，每个目标语言单独采用一个解码器，每个解码器都有自己的注意力机制，但是共享同一个编码器。

实验采用欧洲语料库（EuroParl Corpus），源语言为英语，目标语言分别为法语、西班牙语、荷兰语、葡萄牙语。实验结果显示单语到多语的机器翻译效果均高于英语到其他语言之间的单独翻译，在多数语言对上均提高 1 个 BLEU 值以上。

这种方法共享源语言编码器，能够提高资源稀缺语言对翻译质量。不足之处是每个解码器都拥有单独的注意力机制，计算复杂度较高，限制了在大规模语言对上的应用。

4.3.2 多语到单语翻译

多语到单语翻译是源语言有多个，而目标语言只有一个的机器翻译方法。典型工作为 Zoph 和 Knight[58]提出的多语到单语翻译方法。该方法有两个源语言，分别对应一个编码器，在注意力机制上采用了多源语言注意力机制，这是对 Luong 等人[39]提出的局部注意力的改进。在 t 时刻分别从两个源语言得到上下文向量 和 ，同时应用在解码中。

实验采用 WMT 2014 语料，当源语言为法语、德语，目标语言为英语时，相比一对一的翻译，提高了 4.8 个 BLEU 值；当源语言为英语、法语，目标语言为德语时，则提高 1.1 个 BLEU 值。可以看出，源语言之间的差异对该方法影响很大。除此之外，对每个源语言采用单独的注意力机制，计算复杂度较高。

4.3.3 多语到多语翻译

多语到多语翻译是源语言和目标语言均有多个的机器翻译方法，可以实现多种语言之间互译。Firat 等人[59]提出一种多语言神经机器翻译方法，将该任务定义如下：

假设源语言为 $\{X^1, X^2, \ldots$，目标语言为

$\{Y^1, Y^2, \ldots$，N、M 分别为源语言数量和目标语言数量，需要的双语平行句对为 $\{D_1, \ldots$，$L \leq M$。s和 t表示第 l 对平行语料的源语言和目标语言。针对每个双语对，定义特定平行句对的对数似然值：$\mathcal{L}^{s(D_l)}$。因此，在多语到多语机器翻译中，公式 5 重新定义为：

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{l=1}^{L} \mathcal{L}^{s(D_l),t(D_l} \tag{16}$$

按照如上定义，当模型训练好之后，可以实现语料库中任意两种语言对之间的翻译。

实验采用 WMT 2015 语料，英语到法语、捷克语、德语、俄语、芬兰语之间翻译，共 10 个语言对。相比一对一翻译，多语到多语翻译没有明显提高翻译效果。该方法对每个源语言和目标语言都应用单独的编码器和解码器，但是共享一个注意力机制，降低了计算复杂度，使模型参数随着语言数量呈线性增长。

多语言机器翻译面临参数较多问题，Google 提出一种不改变现有神经机器翻译模型条件下，实现多语到多语的翻译方法[60]。该方法在 GNMT[9]系统上实现，不增加额外参数，仅在平行语料的源语言加入标志，用来指明将要翻译成为哪一种目标语言，并将处理后的多语言平行语料结合起来训练。该方法在单语到多语翻译、多语到单语翻译，以及多语到多语翻译上都表现出较好的效果，并且能够实现语料中没有直接平行对应语言之间的翻译。这种方法不改变现有神经机器翻译模型，实现简单且有效，便于大规模的实际应用。

4.4 规模受限词语表问题

神经机器翻译为了加快训练速度，将双语词典、句子长度限制在一定范围之内。比如，词典由语料中频率较高的词语组成，数量通常限定在 3 万至 8 万之间，其余词语统一用 *unk* 符号表示，同时句子长度大都限定在 50 词以内[19]。这种限制加重了未登录词（OOV）问题，同时低频词也难以学到高质量表示。基于上述，以下是神经机器翻译所要解决的基本问题。

4.4.1 未登录词问题

未登录词问题是语料中部分词语超出了词典覆盖范围，导致该词语不能被准确翻译。神经机器翻译在词典大小限定条件下，随着未登录词数量增加，翻译效果随之严重下降[16]。在现实中，语言是动态变化的系统，词语数量很难固定下来。典型的如人名、地名、机构名等命名实体，另外新词、热词也不断被创造出来。因此，未登录词问题是神经机器翻译的基本研究内容，解决方法大致分为以下三类：

（1）间接方式处理未登录词问题。一种是优化神经网络结构，实现大规模翻译词典或者是开放词典，以此解决未登录词问题，这类方法如 4.4.2 节的第一类方法。另外一种是通过减小源语言和目标语言的翻译粒度，比如采用字符、亚词作为翻译基本单位，以此避免未登录词问题，这类方法如 4.2 节所示。两种方法都可以在一定程度上处理未登录词，不足之处是前者对形态变化较多的语言来说并不是一种有效的方法。

（2）通过上下文信息预测未登录词。该方法基本思想为：如果知道了目标语言未登录词对应的源语言词语，可以通过查找词典将源语言对应的词语转化为目标语言翻译词，或者是根据上下文预测未登录词。现有工作多数源于这种思想。

替换法是最基本的处理方法，当生成未登录词时，通过注意力机制找出该词语对应的源语言词语，将对应概率最大的源语言词语复制过来作为目标语言词语[61]；或者是通过其他词对齐方法，比如统计机器翻译词对齐模型，找出对应源语言词语的对应翻译，以此替换出现的未登录词[18]。这种方法简单直观，有一定的效果，但是忽略了语言的复杂变化和一对多的特殊情况。Luong 等人[62]提出了未登录词标注方法处理未登录词问题，该方法利用源语言、目标语言词语的相对位置信息，能够更加准确的处理未登录词。Li 等人[63]提出了“替换-翻译-恢复”（Substitution-Translation-Restoration）模型，在替换阶段，对测试语料中的低频词用相似的词语替换；在翻译和恢复阶段，利用低频词替换后的语料训练得到翻译模型；翻译时低频词被替换掉，并用替换后的词语进行翻译。这三种方法都可以在一定程度上处理未登录词问题，不同之处是前两种方法并不能处理训练语料之外的未登录词，而第三种方法则可以处理该问题。

（3）以字符或亚词作为翻译基本单位。与 4.2

节所述不同的是本节方法通常作为预处理或后处理，对神经机器翻译模型不作改变。工作主要有 Hirschmann 等人[64]提出的组合词切分方法，以及 Sennrich 等人[49]提出的亚词表示方法。这类方法认为低频词和一些类别的词可以通过比词更小的单位来翻译，比如人名、组合词、同源词和外来词等，关键在于选择合适的切分方法将这些词语切分成粒度合适的亚词，通常选用 BPE 编码。该方法仅作为预处理和后处理，不改变神经机器翻译模型，可以较好地处理未登录词问题。不足之处是输入序列和输出序列长度明显增加，计算量也相应增加。

4.4.2 实现大规模翻译词典

大规模翻译词典指采用的词典较大（与通常 3 万至 8 万相比），或者是大小无限制，一般指目标语言词典。训练神经机器翻译模型困难之一在于计算目标语言词语概率，计算量随着词典增大而增加，从公式 3-4 可以看出。为了在神经机器翻译中应用大规模词典，已有解决方案大致可以分为三类。

（1）优化目标语言词语概率计算方法。Jean 等人[18]提出基于重要性采样（Importance Sampling）近似计算方法，训练中模型每次更新时只用一部分词典；翻译时，可以选择使用全部或部分词典。该方法在采用大规模词典时，大小为 50 万，并没有明显增加训练复杂度，不足之处是训练中采用的目标语言词典大小为 3 万，计算复杂度仍然较高。Mi 等人[65]则在训练中使用句子级词典，大小为 3000。该方法针对每个源语言句子，通过基于词、基于短语的统计机器翻译模型获得每个源语言句子所对应的目标语言词语，并加入 2000 个目标语言常用词，以此构建句子级词典。在 WMT 2015 英语到法语翻译上，与前者相比提高了 1 个 BLEU 值。该方法在速度和翻译质量上均具有显著优势。

（2）对未登录词采用字符级建模，将词语级模型和字符级模型结合在一起。Luong 等人[66]提出一种混合模型，主要采用词语级神经机器翻译模型，对源语言未登录词采用基于字符的表示方法，对目标语言未登录词采用单独的字符级未登录词处理模型。这种方法具有词语级模型训练快速的优点，同时避免了字符级模型序列过长的缺点，通过融合两者优势实现开放词典的神经机器翻译。

（3）词典编码方案。即采用编码方法使神经机器翻译能够在词典大小不变条件下处理更多的源语言、目标语言词语。该方法基于以下思想，假如 V 是个较大的词典，包含了语料中所有词语，W 为一个较小的词典，大小在典型神经机器翻译可以处理的范围内，比如 3 万。如果通过一种编码方案实现将$v \in V$映射到$w \in W$，而不会产生冲突，且可逆，那么就可以在不改变已有翻译模型条件下，实现大规模的翻译词典。根据上述思想，Chitnis 等人[67]提出了基于哈夫曼编码的方案，句子中普通词汇不作处理。通过哈夫曼编码将低频词编码为两个伪词序列（Pseudo-words），总词典大小为普通词汇与伪词个数之和。这种方法对翻译模型本身不作修改，不加入任何额外参数，只需要在翻译前后进行预处理和后处理。

4.4.3 实现长句子翻译

神经机器翻译在约 20 词以内的短句子上取得了很好的效果，随着句子长度增加翻译效果会有所降低[17]。由于训练语料中长句子数量不足，同时循环神经网络存在长时记忆问题，以上是导致长句子翻译效果不好的主要原因。对该问题的处理方法大致分为以下两类。

（1）长句切分方法。将长句子切分成能够被快速翻译的分句片段，然后将分句片段翻译结果组合，最终获得完整句子翻译结果。典型为 Pouget-Abadie 等人[68]的工作，该方法在语序相近语言之间效果较好，不足之处是缺少分句片段之间的长距离调序能力。

（2）增强神经机器翻译表达能力或长距离依赖效果，主要有增强注意力机制[22]、添加外部记忆[28]等方法。

分句切分方法和增强神经网络长距离依赖效果是目前两种主要的处理长句子翻译问题的方法。前者比较简单直观，后者则是解决该问题的根本途径。

4.5 融合外部先验知识方法

先验知识是事先准备的单语、双语、标注数据等，可以指导神经机器翻译的学习过程。多数神经机器翻译模型只依赖句子级词语信息，并不能够学习到充分的语言结构知识，如句法、篇章信息等。在神经机器翻译中融合外部先验知识方法大致可以分为以下几类。

4.5.1 融合统计机器翻译方法

利用统计机器翻译提高神经机器翻译效果是融合先验知识的方法之一。He 等人[69]提出了一种

对数线性神经机器翻译（Log-linear NMT）方法，在生成目标语言词语时，额外加入了词语翻译表和语言模型。翻译表可以提高词汇、低频词翻译效果，语言模型可以提高局部翻译流利度。两种翻译模型分别训练，并通过对数线性模型融合。这种方法属于浅层的融合方式，并没有充分利用神经机器翻译的优势。

与浅层融合相对应的是深层融合方法，典型代表为 Wang 等人[70]的工作。基本思想为生成目标语言词语时，由统计机器翻译提供目标语言候选词列表，用来增强目标语言词语生成质量。通过门机制（Gate Mechanism）将候选词列表和神经机器翻译解码器结合在一起，这两个部分可以联合训练，能够同时利用两种翻译模型的优势。

除上述工作，Zhou 等人[71]提出一种基于神经网络的系统融合框架，将神经机器翻译和统计机器翻译的翻译结果输入到该框架中。在解码中，通过多个注意力机制得到不同系统的翻译结果，通过这种方法获得神经机器翻译和统计机器翻译的共同优点。Stahlberg 等人[72]则将统计机器翻译的最小贝叶斯风险（Minimum Bayes-risk）信息融合到神经机器翻译的解码中，在多个语言对上显著提升了翻译质量。

统计机器翻译研究的时间较长，也较为充分。如何在神经机器翻译模型中充分利用统计机器翻译模型的优势来弥补自身的不足，值得更加深入研究。

4.5.2 增加记忆知识库

以离散形式存储的翻译规则、双语词典等是很重要的翻译知识。将这些翻译知识存储在外部记忆里用于神经机器翻译，是一种有效的处理未登录词、命名实体翻译、术语翻译的方法。

Tang 等人[73]将双语短语对存储在短语记忆（Phrase Memory）里，并用于神经机器翻译。该方法在源语言编码状态 h_i 上加入标记，用于定位和判断不同短语，标记后的编码状态为$\hat{h_i} = [h_i, tag_i]$，***tag*** 为标记向量，表示源语言不同短语。在解码时，能够以词语模式和短语模式生成目标语言，不同模式可以动态控制。该方法主要贡献是在神经机器翻译中集成了短语记忆，改变了解码时一次只能生成一个词语的模式，使之能够一次生成一个词语或者是一个短语。这种方法对命名实体、术语以及未登录词翻译有很大帮助。不足之处是短语记忆里存储的翻译知识仅支持一对一固定翻译。

与上述工作思想类似，Feng 等人[74]提出一种基于记忆的神经机器翻译模型，记忆里存储低频词的翻译规则，以此提高神经机器翻译对低频词、未登录词的处理效果。Wang 等人[75]则将短语记忆集成到神经机器翻译中，用来提高神经机器翻译对短语的翻译质量。

从整体来看，已有工作都没有很充分、高效的将知识库集成到神经机器翻译模型中。

4.5.3 融合双语词典方法

双语词典是很重要的翻译资源，如何充分利用这些已有翻译资源是神经机器翻译迈向实用化的必要步骤。

这方面工作主要有 Arthur 等人[76]提出的将双语词典融入神经机器翻译的方法。首先，通过统计机器翻译词对齐或者是双语词典，获取词语翻译概率；然后，通过注意力机制，将概率化翻译词汇转化为目标语言词语预测概率，可以作为额外先验知识提高目标语言词语生成质量；最后，将这些概率作为偏置，或者是通过插值法，融合到目标语言生成中。这种方法将双语词汇翻译概率作为一种先验知识，提高了实词翻译效果。不足之处是该方法并没有考虑训练语料、词典之外的未登录词。

4.5.4 融合语言学知识

语言学知识可以提高统计机器翻译及其他自然语言处理任务效果[21]。例如，词干提取可以使同一个词的不同形态都归一为一个表示，有利于减少数据稀疏问题。另外，词性标注、句法依存标注信息可在一定程度上提高翻译效果。

通过编码器、解码器融合更多语言信息是利用语言学知识的一种方式。Sennricn 等人[77]将编码器表示为特征组合，即将不同的特征向量拼接在一起。该方法融合特征有词条目特征（Lemma）、亚词标记特征（Subword Tags）、形态特征、词性标注和依存标记特征等。另外，Chen 等人[78]将源语言的依存特征融合到神经机器翻译中；Li 等人[79]以及 Bastings 等人[80]在编码器融合源语言的句法信息；Wu 等[81]人则利用依存树信息增强源语言词语的全局依赖效果；Chen 等人[82]同时在编码器和解码器融合源语言和目标语言句法信息。这类方法扩展了编码器和解码器的结构，并融合更多的语言学特征提高了源语言表示质量，同时也提高了目标语言生成质量。

此外，Niehues 和 Cho[83]采用多任务学习方法将词性标注特征和命名实体特征融合到神经机器翻译中；Zhang 等人[84]提出一种通用的对数线性框架，将先验知识（双语词典、短语表、覆盖惩罚等）集成到神经机器翻译中。

句法树（Syntactic Trees）包含丰富的语言结构信息，将神经机器翻译模型从序列到序列扩展至基于树的形式是研究的热点之一。Eriguchi 等人[85]提出树到序列（Tree-to-Sequence）神经机器翻译模型，采用一个基于树的编码器，自底向上获得源语言句子的短语结构信息。相当于源语言有两个编码器，一个对词语序列信息编码，另外一个对句法结构信息编码，通过注意力机制将两种编码信息融合，在解码时可以同时考虑两种结构形式的信息。Aharoni 和 Goldberg[86]提出一种序列到树（Sequence –to-Tree）神经机器翻译模型，目标语言以线性树（Linearized Trees）形式生成，能够保持序列到序列模型的优点，同时也增强了目标语言的句法结构信息；Wu 等人[87]提出序列到依存（Sequence-to-Dependency）神经机器翻译模型，能够同时对目标语言的词语序列，以及词语之间的依存关系建模，以此提高目标语言的生成质量。以上模型扩展了序列到序列神经机器翻译模型，同时也能够利用更多的语言结构知识指导翻译过程。

4.5.5 融合单语语料方法

单语语料是一种非常重要的资源，具有数量大、获取方便的优势。在统计机器翻译中，大规模目标语言单语语料可以提供优质的语言模型，对提高翻译流利度起着很重要作用。在神经机器翻译中可以利用的单语语料主要分为目标语言单语语料和源语言单语语料。

目标语言单语语料应用之一是语言模型，Gulcehre 等人[88]提出一种利用大规模单语语料提高神经机器翻译效果的方法。采用单语语料训练神经网络语言模型，将之集成到神经机器翻译中，集成方法分为浅层集成和深层集成。浅层集成方法在解码时，把语言模型作为一种特征用来生成候选词；深层集成方法将神经机器翻译模型、语言模型的隐藏状态连接在一起，通过控制机制动态平衡两种模型对解码的影响，在解码时可以捕捉到语言模型信息。这两种集成方法均可以提高翻译效果，其中深层集成方法效果更为明显。此外，Domhan 和 Hieber[89]则提出采用多任务学习方法，将神经机器翻译模型和目标语言的语言模型联合训练，以此利用大规模目标语言单语语料。

目标语言单语语料的另一使用方法是 Sennrich 等人[90]提出的训练数据构造方法：回翻译（Back-translation）方法。利用目标语言单语语料构造伪双语数据，并加入到训练语料。这种融合方法对神经机器翻译模型不作改变，方法简单有效，虽然在一定程度上提高了翻译效果，但是效果提升取决于构造数据的质量。

以上研究都是利用目标语言单语语料，Zhang 和 Zong[91]提出了将源语言单语语料应用到神经机器翻译的方法。实现方式有两种，第一种方法同样采用了构造数据思想，在构造方式上通过自学习（Self-learning）方法扩大双语训练语料规模；另外一种方法通过多任务学习增强编码器对源语言的表示质量。这两种方法均能够大幅提升翻译效果。不足之处是源语言单语语料的数量和题材会对翻译模型性能产生影响。

同时利用源语言和目标语言单语语料，主要有 Cheng 等人[26]提出的半监督学习方法。基本思想是将自编码（Autoencoder）引入源语言到目标语言翻译模型和目标语言到源语言翻译模型，通过半监督方法训练双向神经机器翻译，以此利用源语言和目标语言单语语料提高翻译效果。这种方法显著优势是可以同时利用源语言和目标语言的单语语料，不足之处是对单语语料中的未登录词没有处理能力。此外，Ramachandran 等人[92]提出一种更为简单的方法，将序列到序列模型看作为两个语言模型，通过大规模单语语料分别训练源语言和目标语言的语言模型；神经机器翻译模型的编码器和解码器参数分别由两个语言模型参数初始化；然后利用双语平行语料训练，训练过程中语言模型参数同时调整。

综上所述，大规模单语语料是很重要的资源，如何有效利用这些单语数据是神经机器翻译的重要研究方向。

4.6 资源稀缺条件下的神经机器翻译

神经机器翻译在大规模平行语料条件下取得了显著的效果。但是，在一些资源稀缺语言或领域限定的翻译任务上，平行语料规模相对较少，翻译效果会严重降低[39]。因此，研究资源稀缺条件下的神经机器翻译具有很高实用价值。

融合更多的外部知识是提高资源稀缺语言神经机器翻译效果的方法之一。比如，融合双语词典、融合单语语料、多语言神经机器翻译、多任务学习

等。这些方法本质上是融合了更多的外部知识，提高了神经机器翻译对词语语义、双语词语对应关系的建模能力。

扩充平行语料数量是提高资源稀缺语言神经机器翻译质量的有效方法。如利用回翻译方法[90]快速构建平行语料；通过复制方法构造伪平行语料[93]，即复制目标语言单语句子，作为对应的源语言句子；此外，针对低频词的数据自动增强（Data Augmentation）方法[94]，同样是一种有效的扩充平行语料的方法。

通过迁移学习将资源丰富语言的神经机器翻译模型参数迁移到资源稀缺语言的模型上同样是解决该问题的方法。基于以上思想，Zoph 等人[95]把训练好的语料资源丰富语言对（法语到英语）的模型作为父模型，把语料资源稀缺语言对（西班牙语到英语）作为子模型。通过迁移学习把父模型的一些参数迁移到子模型，方法分为参数初始化和约束训练（Constrain Training）。参数初始化为子模型的一些参数由父模型相同参数初始化，如目标语言词向量。约束训练指子模型的一些参数在训练过程中是固定的。在实验中，该方法显著提升了资源稀缺语言之间的翻译效果。不足之处是在父模型与子模型语言结构相近，并且语料题材相似时，对资源稀缺语言翻译效果提升更为明显。

采用零资源（Zero-resource）神经机器翻译实现资源稀缺语言之间的翻译，同样可以有效处理资源不足问题，通常采用枢轴语言(Pivot language)方法实现。比如 A、B、C 三种语言，A 到 C 之间没有直接的平行语料，但是 A 到 B，B 到 C 之间有较多的平行语料，那么可以将 B 作为枢轴语言，实现 A 到 C 的翻译。在这类工作中，一类是利用隐式的枢轴语言，即不明确指明枢轴语言，如 Google 提出的多语言神经机器翻译方法[60]；另外一类是利用显式枢轴语言，如 Chen 等人[96]提出的“老师—学生”框架；Zheng 等人[97]提出的最大期望似然估计（Maximum Expected Likelihood Estimation）方法；以及 Cheng 等人[98]提出的联合训练方法等。第一类方法需要的参数量少，但是效果较弱；第二类方法参数量大，却能够显著提高翻译效果。

资源稀缺条件下的机器翻译是很多语言面临的现实问题。神经机器翻译的语义表示方法以及方便的参数共享机制为该问题研究提供了便利条件，同时也提供了新的问题解决思路。

4.7 针对评价指标的训练方法

神经机器翻译模型大都通过最大似然估计进行词语级优化，存在以下问题[99]：（1）翻译模型评价面向训练数据，而不是翻译评价标准；（2）损失函数定义在词语级，而不是句子级；（3）在训练中基于训练语料上下文生成目标语言词语，在测试中是基于模型预测上下文生成目标语言词语，而测试过程中的上下文可能存在错误，在后续解码时会被快速放大[100]。

针对神经机器翻译存在的训练问题，Shen 等人[101]将统计机器翻译的最小错误率训练方法（Minimum Risk Training, MRT）引入神经机器翻译。假定 $(x^{(s)},$ 是训练语料中第 s 个句对，y 是模型预测结果。定义损失函数 $\Delta($y计算预测结果 y 和标准翻译 的差异，损失函数可以是机器翻译评测标准，如 BLEU、NIST 等。这种训练方法直接将模型优化与具体任务的评价标准结合在一起，可以应用于任何架构和任意损失函数。

针对循环神经网络训练和测试中存在的问题，Bahdanau 等人[102]提出了判别（Critic）神经网络；Ranzato 等人[99]提出利用增强学习针对评价指标优化的方法；Wisemen 和 Rush[103]则在训练过程中引入序列级损失函数；Norouzi 等人[104]提出了激励增加最大似然算法（Reward augmented Maximum Likelihood，RML）。这些解决序列到序列学习模型问题的方法，同样在神经机器翻译中适用。

综上所述，研究面向翻译质量评价标准的训练方法，是未来神经机器翻译面临的重要挑战。

4.8 新模型与新架构

神经机器翻译相关研究发展迅速，除了传统神经机器翻译模型之外，学者们提出了一些新模型、新架构，主要有以下几种。

4.8.1 多模态神经机器翻译

多模态神经机器翻译利用的资源不限于文本，目前研究主要集中在利用图像信息提高神经机器翻译效果[105-108]。这类方法通常采用两个编码器，一个编码器对文本信息编码，与普通的神经机器翻译相同；另外一个编码器对图像信息编码。在解码时，通过注意力机制将不同模态的信息应用在翻译中。目前工作利用的多模态数据较为单一，有待深

入研究。

4.8.2 非循环神经网络神经机器翻译模型

神经机器翻译模型多数由循环神经网络实现，由于该模型的时序依赖特点，很难并行处理，因而训练和解码速度较慢。Gehring 等人[109]提出了完全基于卷积神经网络的序列到序列模型，相比传统的神经机器翻译模型，速度提升约 10 倍，且翻译质量也有较大提高。Vaswani 等人[110]则抛弃了循环神经网络和卷积神经网络，完全采用注意力机制实现一种序列转导模型（Sequence Transduction），该模型具有很强的并行能力，同时也提高了翻译质量。

4.8.3 新的学习范式

目前，一些学者尝试在神经机器翻译中应用全新的学习范式，如通过对偶学习（Dual Learning）显著降低了平行语料使用量[111]；通过强化学习（Reinforcement Learning）将人工反馈结果应用在神经机器翻译中[112]；Yang 等人[113]以及 Wu 等人[114]分别独立的将生成对抗网络（Generative Adversarial Networks）应用在神经机器翻译中，显著提升了翻译效果。这些探索性工作为神经机器翻译研究提供了全新的视角。

4.9 不同的模型和系统对比分析

神经机器翻译取得了一系列重要研究成果，形成了多个不同的翻译模型，并在一些商业系统上得到应用。本节针对这些不同的模型和系统作简要的对比和分析。

神经机器翻译模型大致可以分为以下几个类别，主要不同之处如表 2 所示。

表 2 神经机器翻译模型对比

翻译模型	翻译单位	注意力	词典限制
经典神经机器翻译	词语	无	限制
注意力神经机器翻译	词语	有	限制
字符级神经机器翻译	字符/亚词	有	不限制
多语言神经机器翻译	词语/亚词	有	可以不限制

经典神经机器翻译模型提出的时间最早，将源语言句子表示成一个固定的向量，解码时依靠该向量生成目标语言词语。这种模型架构简单仅能接近或达到短语统计机器翻译方法[16]。

基于注意力神经机器翻译主要特点是加入了注意力机制，在解码中动态生成源语言相关信息，提高了模型的表达能力和长距离依赖效果，是目前主流方法。

字符级神经机器翻译主要特点是以字符、亚词作为翻译基本单位，能够在一定程度上避免未登录词问题，因此对翻译词典大小不作限制。

多语言神经机器翻译主要特点是将一对一的翻译模型扩展成一对多、多对一或多对多的翻译模型。目前主要针对西方拼音文字之间的翻译，翻译基本单位可以采用词语、亚词等，当采用亚词时，能够实现开放词典翻译。

为了对比不同的翻译系统，我们采用 2009 年中文信息学会主办的机器翻译评测英汉翻译测试集，对不同翻译系统进行评测，结果如表 3 所示。

表 3 不同翻译系统性能对比

翻译系统	BLEU4	BLEU5	BLEU6	BLEU7
百度翻译	48.96	41.63	35.36	29.97
Google 翻译	50.18	42.94	36.65	31.25
小牛翻译	51.67	44.30	37.81	32.18
搜狗翻译	60.72	53.74	47.47	41.88

从表 3 可以看出，所有翻译系统译文质量达到了较高的水平。以上系统主要采用神经机器翻译模型，翻译结果很可能融合了多种模型，并且训练数据规模不公开，有可能包含本文所采用的测试集，因此评测结果只能作为近似对比。

从译文分析中可以看到神经机器翻译共同优点是译文比较流畅，同时都存在命名实体翻译质量较差问题，且翻译不充分问题仍然存在。

5 基于神经网络的机器翻译评测

系统评测是比较不同机器翻译系统性能好坏的重要方法，分为主观评测和客观评测。主观评测采用人工主观判断对翻译系统译文打分，评判标准为流利度和忠实度等[21]。主观评测方法虽然质量很高，但是不免存在评分不一致问题，并且需要经验丰富的专家才能胜任，评测代价很高。客观评测方法采用一定的模型对译文打分，常用评测方法有 BLEU、NIST 等[21]。客观评测具有速度快、效率高的优点，但是评测结果并不能完全反映译文质量好坏，仍然存在不足之处。

机器翻译不仅仅是一种语言字符串到另外一种语言字符串的转换，而应该是一种语言所表达的语义到另外一种语言的同等语义的完全表达。传统评测方法主要依据字符串、词典的匹配程度，另外

一些评测方法融合了语言知识、知识库等，提高了译文评测质量。这些融合语言知识的综合方法本质上是提高了词义、语义知识的评测比重，更为符合译文评测的实质。词向量可以表达丰富的语义和语言结构信息，是一种相对理想的语义表示方法。用神经网络抽象出有效特征，或者是采用神经网络模型进行翻译评测，成为当前热点研究[115]。

机器翻译译文评测可以看作在候选译文中，区分质量好的译文和质量差的译文。Guzman 等人[116]提出了一种基于神经网络的机器翻译评测方法，依据参考译文，从翻译系统生成的 2 个候选翻译选项中找出最好的翻译。方法如下：设 t_1，t_2 是两个候选翻译，r 是对应的参考翻译。可以用分类器从两个候选翻译中找出最好的翻译，如下所示，y 表示分类结果。

$$y=\begin{cases}1 & \text{给定}r,t_1\text{优}\\ 0 & \text{给定}r,t_2\text{优}\end{cases} \tag{17}$$

定义上述分类任务之后，用前馈神经网络建模。当给定待评测译文和参考译文对(t_1, t_2, r)，将其映射到维数固定的向量$(\mathrm{x}_{t1}, \mathrm{x}_{t2}, \mathrm{x}_r)$，并融入句法、语义信息，然后输入上述模型，并分类。

这种评测方法可以融合参考译文的句法、语义信息，以及两个候选翻译特征。用向量表示上述特征，并通过神经网络建模。不足之处是只能在已有的翻译候选结果中找出翻译质量较高的候选翻译，并不能给出具体评测分值。

Gupta 等人[117]提出了一种简单有效的基于循环神经网络的机器翻译评测方法，可以对待评测译文给出具体评测分值，方法如下：

h_{ref}，h_{tra} 分别表示参考译文和译文，$\hat{y}$表示两者的相似度，通过循环神经网络计算得到，如下所示：

$$h_{\times}=h_{ref}\odot \tag{18}$$

$$h_{+}=|h_{ref}- \tag{19}$$

$$h_s=\sigma(W^{(\times)}h_{\times}+W^{+}h_{+}+ \tag{20}$$

$$\hat{p}_\theta=softmax(W^{(p)}h_s+ \tag{21}$$

$$\hat{y}=1 \tag{22}$$

$\hat{p}_\theta$是计算得到的概率分布向量，$r^T=[1\ 2\ldots K]$。

该方法可以融合不同层次的语义特征，利用外部资源较少，在 WMT 2014 评测任务的 5 个语言对上取得了最好的成绩。

设计有效特征在机器翻译评测中起着重要作用。采用的特征包括从简单、语言无关的基本特征，到基于语言结构的高级特征。这种人工设计的特征具有领域相关性，在不同的数据集和语言上的应用效果会发生变化，并且大都忽略了上下文信息。鉴于上述问题，Shah 等人[115]提出了一些采用神经网络训练得到的特征，包括连续空间语言模型特征，大规模单语语料训练得到的词向量特征，通过词对齐和词语表示计算得到的目标语言词语与源语言词语相似度等特征。这些特征都是通过无监督方法训练得到，简单有效，适应性强。

翻译评测对机器翻译研究起着引导作用，是机器翻译重要研究方向。如何结合神经网络的优势构建新的评测方法，使自动评测结果更为符合人类专家对翻译质量的评价是机器翻译评测的重要目标。

6 未来研究方向

目前，神经机器翻译取得巨大成功，新的研究成果不断涌现出来，可以称作统计机器翻译之后一种全新的机器翻译方法。严格来讲，从 2014 年开始，神经机器翻译得到人们的广泛关注[15-19]，随后大量相关成果发表出来。由于研究时间较短，该翻译模型仍然存在许多值得更加深入探索的问题，以下几点有可能成为未来研究集中方向。

（1）提高语言学解释性：基于编码器解码器的神经机器翻译，实现了源语言到目标语言的直接翻译，但是翻译过程很难得到充分的语言学解释。已有工作证明，可以从词语级神经机器翻译编码器中抽取出隐含的句法结构信息[118]，以及在一定程度上对神经机器翻译的翻译过程进行解释和分析[119]。从神经机器翻译模型中抽取出相应的语言学知识来解释翻译过程，以此改进翻译模型，是神经机器翻译未来重要的研究方向。

（2）融合外部先验知识：以离散符号表示的外部资源，如句法标注、词性标注、双语词典等是非常重要的先验知识，在神经机器翻译中难以得到充分利用。融合更加丰富的先验知识是神经机器翻译重要研究内容，也是提高翻译效果的重要方法，有待深入研究。

（3）基于句法的神经机器翻译：神经机器翻

译大都是词语级的序列到序列模型，所包含的句法信息较少。句法是重要的关于句子结构的理论，将序列到序列翻译模型扩展至基于句法的翻译模型，如树到序列[85]、序列到树[86]、树到树等，是神经机器翻译模型架构创新的重要体现。

（4）多语言机器翻译：连续空间表示法是有效的多语语义表示方法[13]，注意力机制经实验证明能够在不同语言之间共享[59]，这些为多语言机器翻译研究提供了良好的基础。在多语平行语料，或者多语可比语料基础上研究基于神经网络的多语言机器翻译，不仅具有学术价值同样具有很高的实用价值，是未来重要的发展方向。

（5）多模态翻译：神经网络能够以统一的形式对文字、图像、语音等不同模态数据进行表示。目前，文字与图像之间实现端到端的直接翻译[120]，并且图像信息也被应用到神经机器翻译[107]。高效利用文字本身以外的信息，如语音、图像、位置场景等，以此构建多模态翻译是机器翻译真正实用化的必经之路。

7 小 结

实现不同语言之间的无障碍交流是计算机发明初期就开始追逐的梦想。经过了 60 多年的发展，从基于规则的机器翻译到基于统计的机器翻译，以至当前的神经机器翻译，从整体来看是在不断地降低人们对翻译过程的干预程度。从翻译效果上看，在同等条件下神经机器翻译的翻译质量更好，同时仍有很大的提升空间。对于统计机器翻译难以有效处理的多语言机器翻译、长距离调序问题、模型迁移问题等，也能够高效处理。神经网络为机器翻译研究打开了广阔的视野。

神经机器翻译代表了一种全新的机器翻译模型，在部分语言上逐渐呈现出全面超越统计机器翻译趋势。虽然该方法在模型架构、训练算法、可解释性等方面存在不足之处，但是必将成为未来机器翻译的发展方向。

参 考 文 献

[1] Jiao Li-Cheng, Yang Shu-Yuan, Liu Fang, et al. Seventy years beyond neural networks: retrospect and prospect. Chinese Journal of Computers, 2016, 39(8):1697–1716. (in Chinese)
（焦李成，杨淑媛，刘芳等.神经网络七十年：回顾与展望. 计算机学报，2016, 39(8):1697–1716）

[2] Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18:1527–1554

[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks//Proceedings of the Neural Information Processing Systems (NIPS 2012). Lake Tahoe, USA, 2012:1097–1105

[4] Geoffrey Hinton, Li Deng, Dong Yu. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6):82–97

[5] Ronan Collobert, Jason Weston, Leon Bottou, et al. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12(1):2493–2537

[6] Marcin Junczys-Dowmunt, Tomasz Dwojak, Hieu Hoang. Is neural machine translation ready for deployment? A case study on 30 translation directions. arXiv preprint/1610.01108v2, 2016

[7] Rico Sennrich, Barry Haddow, Alexandra Birch. Edinburgh neural machine translation systems for WMT 16//Proceedings of the First Conference on Machine Translation. Berlin, Germany, 2016:371–376

[8] Jie Zhou, Ying Cao, Xuguang Wang, et al. Deep recurrent models with fast-forward connections for neural machine translation. Transactions of the Association for Computational Linguistics, 2016, 4:371–383

[9] Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint/1609.08144v1, 2016

[10] Josep Crego, Jungi Kim, Guillaume Klein, et al. SYSTRAN's pure neural machine translation systems. arXiv preprint/1610.05540v1, 2016

[11] Ramnn P. Neco, Mikel L. Forcada. Asynchronous translations with recurrent neural nets//Proceedings of International Conference on Neural Networks. Houston, USA, 1997:2535–2540

[12] M.A. Castano, F. Casacuberta. A connectionist approach to machine translation//Proceedings of Fifth European Conference on Speech Communication and Technology. Rhodes, Greece, 1997:1–4

[13] Jiajun Zhang, Chengqing Zong. Deep neural networks in machine translation: an overview. Intelligent Systems IEEE, 2015, 30(5):16–25

[14] Nal Kalchbrenner, Phil Blunsom. Recurrent continuous translation models//Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Seattle, USA, 2013:1700–1709

[15] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to sequence learning with neural networks//Proceedings of the Neural Information Processing Systems (NIPS 2014). Montreal, Canada, 2014:3104–3112

[16] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. arXiv preprint/1406.1078v2, 2014

[17] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, et al. On the properties of neural machine translation: encoder-decoder approaches//Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar, 2014: 103–111

[18] Sebastien Jean, Kyunghyun Cho, Yoshua Bengio. On using very large target vocabulary for neural machine translation//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015). Beijing, China, 2015:1–10

[19] Sebastien Jean, Orhan Firat, Kyunghyun Cho, et al. Montreal neural machine translation systems for WMT'15//Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, Portugal, 2015:134–140

[20] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, et al. Neural versus phrase-based machine translation quality//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:257–267

[21] Zong Cheng-Qing. Statistical machine ttranslation. Second Edition. Beijing: Tsinghua University Press, 2013 (in Chinese)
（宗成庆. 统计自然语言处理. 第 2 版. 北京：清华大学出版社，2013）

[22] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint/1409.0473v6, 2014

[23] Zhaopeng Tu, Zhengdong Lu, Liu Yang, et al. Modeling coverage for neural machine translation//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:76–85

[24] Gregory Druck, Kuzman Ganchev, and Joao Graca. Rich prior knowledge in learning for NLP//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011). Portland, USA, 2011:1–57

[25] Zhaopeng Tu, Yang Liu, Lifeng Shang, et al. Neural machine translation with reconstruction//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017). San Francisco, USA, 2017:3097–3103

[26] Yong Cheng, Wei Xu, Zhongjun He, et al. Semi-supervised learning for neural machine translation//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:1965–1974

[27] Biao Zhang, Deyi Xiong, Jinsong Su. Variational neural machine translation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:521–530

[28] Mingxuan Wang, Zhengdong Lu, Hang Li, et al. Memory-enhanced decoder for neural machine translation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:278–286

[29] Jeffrey L Elman. Finding structure in time. Cognitive Science, 1990, 14(2):179–211

[30] Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep learning. Cambridge, USA: MIT Press, 2015

[31] Yoshua Bengio, Patrice Simard, Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 1994, 5(2):157–66

[32] Sepp Hochreiter, Jurgen Schmidhuber. Long short-term memory. Neural Computation, 1997, 9(8):1735–1780

[33] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint/1412.3555v1, 2014

[34] Richard Socher, Eric H. Huang, Jeffrey Pennington. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection// Proceedings of the Neural Information Processing Systems (NIPS 2011). Granada, Spain, 2011:801–809

[35] Alex Graves, Greg Wayne, Malcolm Reynolds, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 2016, 538:471–476

[36] Alex Graves, Greg Wayne, Ivo Danihelka. Neural turing machines. arXiv preprint/1410.5401v2, 2014

[37] Jason Weston, Sumit Chopra, Antoine Bordes. Memory networks. arXiv preprint/1410.3916v11, 2014

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, et al. Show, attend and tell: neural image caption generation with visual attention. arXiv preprint/1502.03044, 2015

[39] Minh-Thang Luong, Hieu Pham, Christopher D. Manning. Effective approaches to attention-based neural machine translation// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal, 2015:1412–1421

[40] Lemao Liu, Masao Utiyama, Andrew Finch, et al. Neural machine

translation with supervised attention//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. Osaka, Japan, 2016:3093–3102

[41] Yang Liu, Maosong Sun. Contrastive unsupervised word alignment with non-local features//Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015). Austin, USA, 2015:2295–2301

[42] Yong Cheng, Shiqi Shen, Zhongjun He, et al. Agreement-based joint training for bidirectional attention-based neural machine translation//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2016). New York, USA, 2016:2761–2767

[43] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, et al. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 1993, 19(2):263–311

[44] Shi Feng, Shujie Liu, Mu Li, et al. Implicit distortion and fertility models for attention-based encoder-decoder NMT model. arXiv preprint/1601.03317v3, 2016

[45] Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, et al. Incorporating structural alignment biases into an attentional neural translation model. arXiv preprint/1601.01085v1, 2016

[46] Jinchao Zhang, Mingxuan Wang, Qun Liu, et al. Incorporating word reordering knowledge into attention-based neural machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada , 2017:1524–1534

[47] Zhaopeng Tu, Yang Liu, Zhengdong Lu, et al. Context gates for neural machine translation. Transactions of the Association for Computational Linguistics , 2017, 5:87–99

[48] Yoon Kim, Yacine Jernite, David Sontag, et al. Character-aware neural language models. arXiv preprint/1508.06615v4, 2015

[49] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural machine translation of rare words with subword units. arXiv preprint/1508.07909v3, 2015

[50] Junyoung Chung, Kyunghyun Cho, Yoshua Bengio. A character-level decoder without explicit segmentation for neural machine translation//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:1693–1703

[51] Marta R. Costa-jussa, Jose A. R. Fonollosa. Character-based neural machine translation. arXiv preprint/1603.00810v2, 2016

[52] Jinsong Su, Zhixing Tan, Deyi Xiong, et al. Lattice-based recurrent neural network encoders for neural machine translation//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017). San Francisco, USA, 2017:3302–3308

[53] Zhen Yang, Wei Chen, Feng Wang, et al. A character-aware encoder for neural machine translation//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. Osaka, Japan, 2016:3063–3070

[54] Wang Ling, Isabel Trancoso, Chris Dyer, et al. Character-based neural machine translation. arXiv preprint/1511.04586v1, 2015

[55] Jason Lee, Kyunghyun Cho, Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. arXiv preprint/1610.03017v1, 2016

[56] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever , et al. Multi-task sequence to sequence learning. arXiv preprint/1511.06114v4, 2015

[57] Daxiang Dong, Hua Wu, Wei He, et al. Multi-task learning for multiple language translation//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015). Beijing, China, 2015:1723–1732

[58] Barret Zoph, Kevin Knight. Multi-source neural translation. arXiv preprint/1601.00710, 2016

[59] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism//Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016). San Diego, USA, 2016:866–875

[60] Melvin Johnson, Mike Schuster, Quoc V. Le, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 2017, 5:339-351

[61] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, et al. Pointing the unknown words//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:140–149

[62] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, et al. Addressing the rare word problem in neural machine translation//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015). Beijing, China, 2015:11–19

[63] Xiaoqing Li, Jiajun Zhang, Chengqing Zong. Towards zero unknown word in neural machine translation//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2016). New York, USA, 2016:2852–2858

[64] Fabian Hirschmann, Jinseok Nam, and Johannes Furnkranz. What makes word-level neural machine translation hard:a case study on english-german translation//Proceedings of COLING 2016, the 26th

International Conference on Computational Linguistics. Osaka, Japan, 2016:3199–3208

[65] Haitao Mi, ZhiguoWang, Abe Ittycheriah. Vocabulary manipulation for neural machine translation//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:124–129

[66] Minh-Thang Luong, Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:1054–1063

[67] Rohan Chitnis, John DeNero. Variable-length word encodings for neural translation models//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal, 2015:2088–2093

[68] Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merrienboer, et al. Overcoming the curse of sentence length for neural machine translation using automatic segmentation//Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8). Doha, Qatar, 2014:78–85

[69] Wei He, Zhongjun He, Hua Wu, et al. Improved neural machine translation with SMT features//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016). Phoenix, USA, 2016:151–157

[70] Xing Wang, Zhengdong Lu, Zhaopeng Tu, et al. Neural machine translation advised by statistical machine translation//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017). San Francisco, USA, 2017:3330–3336

[71] Long Zhou, Wenpeng Hu, Jiajun Zhang, et al. Neural system combination for machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:378–384

[72] Felix Stahlberg, Adria de Gispert, Eva Hasler, et al. Neural machine translation by minimising the bayes-risk with respect to syntactic translation lattices//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017). Valencia, Spain, 2017:362–368

[73] Yaohua Tang, Fandong Meng, Zhengdong Lu, et al. Neural machine translation with external phrase memory. arXiv preprint/1606.01792v1, 2016

[74] Yang Feng, Shiyue Zhang, Andi Zhang, et al. Memory-augmented neural machine translation//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:1401–1410

[75] Xing Wang, Zhaopeng Tu, Deyi Xiong, et al. Translating phrases in neural machine translation//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:1432–1442

[76] Philip Arthur, Graham Neubig, Satoshi Nakamura. Incorporating discrete translation lexicons into neural machine translation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:1557–1567

[77] Rico Sennrich, Barry Haddow. Linguistic input features improve neural machine translation//Proceedings of the First Conference on Machine Translation. Berlin, Germany, 2016:83–91

[78] Kehai Chen, Rui Wang, Masao Utiyama, et al. Neural machine translation with source dependency representation//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:2836–3842

[79] Junhui Li, Deyi Xiong, Zhaopeng Tu, et al. Modeling source syntax for neural machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:688–697

[80] Joost Bastings, Ivan Titov, Wilker Aziz, et al. Graph convolutional encoders for syntax-aware neural machine translation//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:1947–1957

[81] Shuangzhi Wu, Ming Zhou and Dongdong Zhang. Improved neural machine translation with source syntax//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017). Melbourne, Australia, 2017:4179–4185

[82] Huadong Chen, Shujian Huang, David Chiang, et al. Improved neural machine translation with a syntax-aware encoder and decoder//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:1936–1945

[83] Jan Niehues and Eunah Cho. Exploiting linguistic resources for neural machine translation using multi-task learning//Proceedings of the Conference on Machine Translation (WMT 2017). Copenhagen, Denmark, 2017:80–89

[84] Jiacheng Zhang, Yang Liu, Huanbo Luan, et al. Prior knowledge integration for neural machine translation using posterior regularization//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:1514–1523

[85] Akiko Eriguchi, Kazuma Hashimoto, Yoshimasa Tsuruoka. Tree-to-Sequence attentional neural machine translation//Proceedings

of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016: 823–833

[86] Roee Aharoni and Yoav Goldberg. Towards string-to-tree neural machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:132–140

[87] Shuangzhi Wu, Dongdong Zhang, Nan Yang, et al. Sequence-to-Dependency neural machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:698–707

[88] Caglar Gulcehre, Orhan Firat, Kelvin Xu, et al. On using monolingual corpora in neural machine translation. arXiv preprint/1503.03535v2, 2015

[89] Tobias Domhan and Felix Hieber. Using target-side monolingual data for neural machine translation through multi-task learning//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:1501–1506

[90] Rico Sennrich, Barry Haddow, Alexandra Birch. Improving neural machine translation models with monolingual data//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:86–96

[91] Jiajun Zhang, Chengqing Zong. Exploiting source-side monolingual data in neural machine translation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:1535–1545

[92] Prajit Ramachandran, Peter J. Liu and Quoc V. Le. Unsupervised pretraining for sequence to sequence learning//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:383–391

[93] Anna Currey, Antonio Barone and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation// Proceedings of the Conference on Machine Translation (WMT 2017). Copenhagen, Denmark, 2017:148–156

[94] Marzieh Fadaee, Arianna Bisazza and Christof Monz. Data augmentation for low-resource neural machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:567–573

[95] Barret Zoph, Deniz Yuret, Jonathan May, et al. Transfer learning for low-resource neural machine translation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:1568–1575

[96] Yun Chen, Yang Liu, Yong Cheng, et al. A teacher-student framework for zero-resource neural machine translation// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:1925–1935

[97] Hao Zheng, Yong Cheng, Yang Liu. Maximum expected likelihood stimation for zero-resource neural machine translation// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017). Melbourne, Australia, 2017:4251–4257

[98] Yong Cheng, Qian Yang, Yang Liu, et al. Joint training for pivot-based neural machine translation//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017). Melbourne, Australia, 2017:3974

[99] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, et al. Sequence level training with recurrent neural networks. arXiv preprint/1511.06732v7, 2015

[100] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, et al. Scheduled sampling for sequence prediction with recurrent neural networks// Proceedings of the Neural Information Processing Systems (NIPS 2015). Montreal, Canada, 2015:1–9

[101] Shiqi Shen, Yong Cheng, Zhongjun He, et al. Minimum risk training for neural machine translation//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:1683–1692

[102] Dzmitry Bahdanau, Philemon Brakel, Ryan Lowe, et al. An actor-critic algorithm for sequence prediction. arXiv preprint/1607.07086v2, 2016

[103] Sam Wiseman, Alexander M. Rush. Sequence-to-Sequence learning as beam-search optimization//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:1296–1302

[104] Mohammad Norouzi, Samy Bengio, Zhifeng Chen, et al. Reward augmented maximum likelihood for neural structured prediction// Proceedings of the Neural Information Processing Systems (NIPS 2016). Barcelona, Spain, 2016:1723–1731

[105] Iacer Calixto, Qun Liu and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:1913–1924

[106] Jean-Benoit Delbrouck and Stephane Dupont. An empirical study on the effectiveness of images in Multimodal Neural Machine Translation//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:921–930

[107] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation//Proceedings of the 2017

Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:1003–1014

[108] Ozan Caglayan, Walid Aransa, Yaxing Wang, et al. Does multimodality help human and machine for translation and image captioning?//Proceedings of the First Conference on Machine Translation (WMT 2016). Berlin, Germany, 2016:627–633

[109] Jonas Gehring, Michael Auli, David Grangier, et al. Convolutional sequence to sequence learning. arXiv preprint/1705.03122v3, 2017

[110] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. arXiv preprint/1706.03762v4, 2017

[111] Di He, Yingce Xia, Tao Qin, et al. Dual learning for machine translation//Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016). Barcelona, Spain, 2016:1–9

[112] Khanh Nguyen, Hal Daume III, and Jordan Boyd-Graber. Reinforcement learning for bandit neural machine translation with simulated human feedback//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:1465–1475

[113] Zhen Yang, Wei Chen, Feng Wang, et al. Improving neural machine translation with conditional sequence generative adversarial nets. arXiv preprint/1703.04887v2, 2017

[114] Lijun Wu, Yingce Xia, Li Zhao, et al. Adversarial neural machine translation. arXiv preprint/1704.06933v3, 2017

[115] Kashif Shah, Varvara Logacheva, Gustavo Henrique Paetzold. SHEF-NN: translation quality estimation with neural networks//Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, Portugal, 2015:342–347

[116] Francisco Guzman, Shafiq Joty, Lluıs Marquez, et al. Pairwise neural machine translation evaluation//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015). Beijing, China, 2015:805–814

[117] Rohit Gupta, Constantin Orasan, Josef van Genabith. ReVal: a simple and effective machine translation evaluation metric based on recurrent neural networks//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal, 2015: 1066–1072

[118] Xing Shi, Inkit Padhi, Kevin Knight. Does string-based neural mt learn source syntax//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:1526–1534

[119] Yanzhuo Ding, Yang Liu, Huanbo Luan, et al. Visualizing and understanding neural machine translation//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017:1150–1159

[120] Scott Reed, Zeynep Akata, Xinchen Yan, et al. Generative adversarial text to image synthesis//Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). New York, USA, 2016:1060–1069

LI Ya-Chao, born in 1986, Ph.D. candidate, lecturer. His research interests include machine translation and natural language processing.

XIONG De-Yi, born in 1979, Ph.D., professor. His research interests include natural language processing, statistical machine translation, and multilingual information access.

ZHANG Min, born in 1970, Ph.D., professor, Ph.D. supervisor. His research interests include machine translation and natural language processing.

Background

Due to success of sequence-to-sequence models, neural machine translation is proposed for machine translation based purely on neural networks, leading to the state-of-the-art machine translation model over various language pairs. Under the background of internationalization, multicultural communication is one of the defining trends of the next few decades. Rapid international development drives growth in the increasingly demand of convenient language translation service. The growing demand for language services is no longer concentrated in a few large businesses and governments, and all kinds of other organizations have increased their use of translation services. Considering the big market, many companies, such as Google, Microsoft, and Baidu have launched different translation services based on statistical machine translation a few years ago. However, after nearly 20 years of development, the statistical machine translation method has entered a bottleneck period. In recent years, academia and industry are exploring smart machine translation solutions.

In the context of the rapid development of deep learning

technology and artificial intelligence solutions, NMT has achieved fruitful results in recent years. In this paper, the authors present a survey on models and techniques of neural machine translation. They explain a simple encoder-decoder model firstly, then describe how an attention mechanism can be incorporated into the encoder-decoder model and the research progress of NMT in detail. The authors also discuss the different models and systems of NMT. Finally, the authors give some concluding remarks on new challenges for and new directions in future research of NMT.

In recent years, the author's group has focus on the related research with machine translation, such as neural machine translation, statistical machine translation, syntax parsing, and multilingual information access. This work is supported by the National Natural Science Foundation of China (61525205, 61432013, and 61403269), the Fundamental Research Funds for the Central Universities of Northwest Minzu University (31920170154 and 31920170153), and the Scientific Research Project of Universities in Gansu (2016B-007).