神经机器翻译

王星 熊德意 张民

苏州大学

1. 引言

神经机器翻译（Neural Machine Translation）是指直接采用神经网络以端到端方式进行翻译建模的机器翻译方法。区别于利用深度学习技术完善传统统计机器翻译中某个模块的方法，神经机器翻译采用一种简单直观的方法完成翻译工作：首先使用一个称为编码器（Encoder）的神经网络将源语言句子编码为一个稠密向量，然后使用一个称为解码器（Decoder）的神经网络从该向量中解码出目标语言句子。上述神经网络模型一般称之为“编码器-解码器”（Encoder-Decoder）结构（见图 1）。

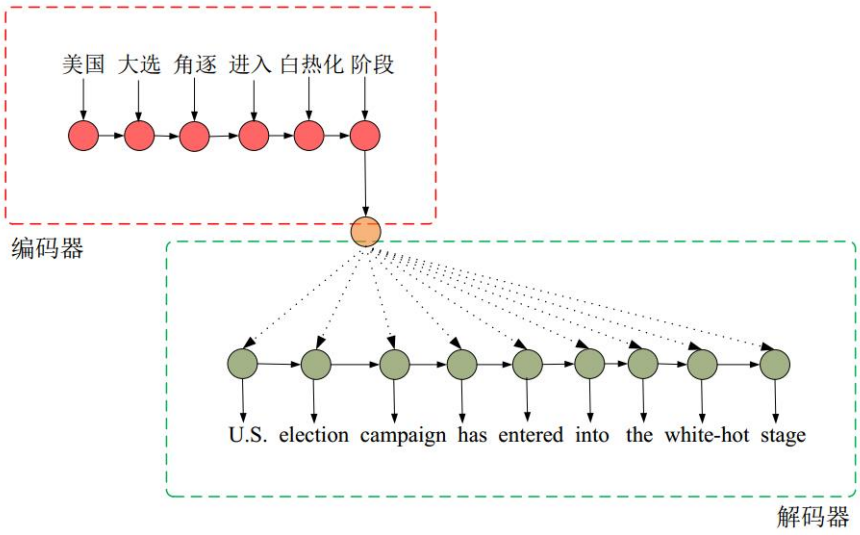

图 1. 神经机器翻译“编码器-解码器”结构

神经机器翻译的建模思想可以追溯至上世纪 90 年代，西班牙阿利坎特大学的 Forcada 和 Ñeco [1]于 1997 年提出“编码器-解码器”结构进行翻译转换工作。在深度学习和分布式词语表示成功应用于自然语言处理领域的背景下，英国牛津大学 Kalchbrenner 和 Blunsom[2]于 2013 年以连续表示为基础提出了神经机器翻译。加拿大蒙特利尔大学 Cho 等人[3]和谷歌公司 Sutskever 等人[4]于 2014 年分别对该方法进行了完善。Bahdanau 等人[5]在 Cho 等人[3]工作的基础上引入注意力机制（attention mechanism），显著提高神经机器翻译模型的翻译性能。由于其结构简单且性能显著，采用注意力机制的神经机器翻译引起全球科研人员的广泛

关注和研究，获得了迅速的发展。

虽然神经机器翻译取得了不错的翻译性能，但是上述研究[2][3][4][5]的实验均是在单个翻译任务（英语-法语）上以自动评价指标 BLEU 为标准进行翻译性能评测。神经机器翻译的性能难免面临更多的质疑：是否在其他语言对上也有类似的实验结果？是否在人工评测中仍会胜出？是否在大规模语料上仍然有效？针对上述问题，一些研究者开展神经机器翻译和传统统计机器翻译的对比工作。Junczys-Dowmunt 等人[6]在联合国平行语料库（United Nations Parallel Corpus v1.0）30 个语言对上开展对比工作。实验表明，以 BLEU 值为评测指标，与传统的基于短语的统计机器翻译相比，神经机器翻译具有压倒性的优势：神经机器翻译在 27 个语言对上超过了基于短语的统计机器翻译，仅在 2 个语言对上以微弱的劣势落败。值得注意的是，神经机器翻译在涉及汉语的翻译任务上比短语系统能够提高 4 至 9 个 BLEU 点，性能提高尤其显著。Bentivogli 等人[7]对口语翻译国际研讨会（International Workshop on Spoken Language Translation，IWSLT）评测任务中英语-德语翻译任务的官方结果进行深入的人工分析和对比。他们发现，相比基于短语的机器翻译，神经机器翻译不仅在人工评测指标上占优，而且能够减少词法错误（morphology errors）、词汇错误（lexical errors）和词序错误（word order errors）。谷歌公司 Wu 等人[8]的实验表明在大规模语料情况下，神经机器翻译在实验所涉及的 6 个语言对的翻译任务上，人工评测结果仍能占优。

与此同时，神经机器翻译在国际机器翻译公开评测中性能也达到或者超出传统统计机器翻译方法。在 2015 年的统计机器翻译研讨会（Workshop on Statistical Machine Translation，WMT）评测任务所发布的官方评测结果中[9]，蒙特利尔大学的神经机器翻译系统取得了不错的成绩：在英语-德语翻译任务上斩获头名，在德语-英语、捷克语-英语、英语-捷克语的翻译任务上取得第三名。在 2015 年的口语翻译国际研讨会评测任务所发布的官方评测结果中[10]，斯坦福大学的神经机器翻译系统在英语-德语翻译任务上夺得头名。

2015 年 5 月，百度翻译发布融合统计和深度学习方法的在线翻译系统，以提升在线翻译质量。2016 年 9 月，谷歌翻译在汉语-英语方向上采用了内部开发的神经翻译系统（Google Neural Machine Translation System）替代其旧版所使用的基于短语的翻译系统（Phrase-Based Machine Translation System）。神经机器翻译

在工业界的应用为其发展壮大进一步奠定了基础。

采用注意力机制的神经机器翻译有哪些特点？神经机器翻译译文存在哪些问题？神经机器翻译的当前研究热点有哪些？神经机器翻译的未来研究趋势在哪里？本文尝试对这几个问题进行阐述。

2. 采用注意力机制的神经机器翻译模型

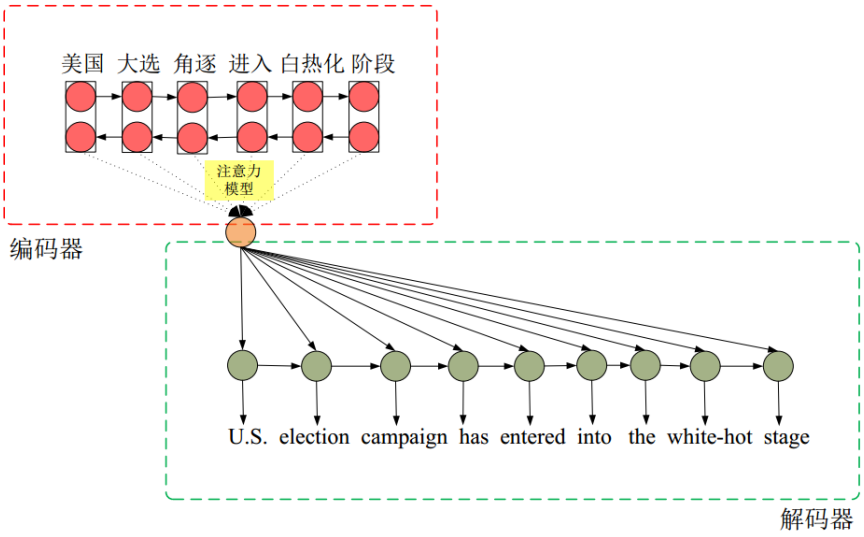

图 2. 采用注意力机制的神经机器翻译“编码器-解码器”结构

本节对当前流行的采用注意力机制的神经机器翻译进行简单介绍。

采用注意力机制“编码器-解码器”结构与普通结构不同之处在于稠密向量（后文称之为上下文向量）的构建。普通的“编码器-解码器”结构采用前向循环神经网络编码器对源语言句子进行编码并将末尾的隐式状态作为上下文向量（见图 1 红色方框“编码器”部分）。而在基于注意力机制的“编码器-解码器”结构采用双向循环神经网络编码器对源语言句子进行编码。解码步骤中，模型通

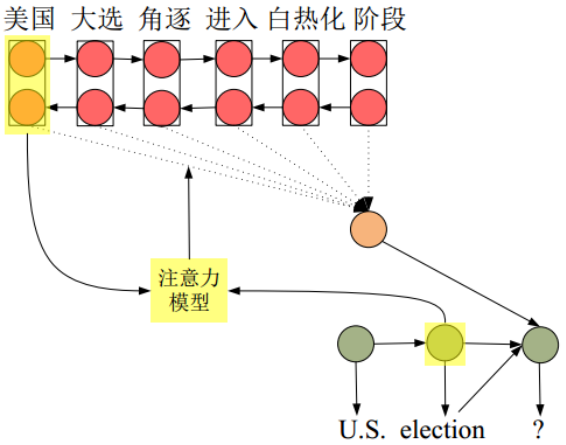

图 3. 采用注意力机制的神经机器翻译的工作流程

过注意力机制选择性地关注源语言句子的不同部分，动态地构建上下文向量（见图 2 红色方框“编码器”部分）。

采用注意力机制的神经机器翻译的工作流程如图 3 所示。给定源语言句子$\mathrm{X}=\{x_1, x_2, \ldots x_n\}$。双向循环神经网络编码器将句子 X 编码为一个源语言隐式状态序列$\mathrm{H}=\{h_1, h_2, \ldots h_n\}$，其中前向循环神经网络顺序读入句子X后产生源语言前向隐式状态序列$\overrightarrow{\mathrm{H}}=\{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots \overrightarrow{h_n}\}$，后向循环神经网络逆序读入句子X后产生源语言后向隐式状态序列$\overleftarrow{\mathrm{H}}=\{\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots \overleftarrow{h_n}\}$。前向和后向隐式状态序列中位置对应的状态序列拼接形成该位置单词的隐式状态$h_k=[\overrightarrow{h_k}; \overleftarrow{h_k}]$。

在解码时刻 t，解码器分别产生该时刻的目标语言隐式状态和目标语言单词。t 时刻目标语言隐式状态s_t由 t-1 时刻目标语言隐式状态s_{t-1}，t-1 时刻解码器所生成的目标语言单词y_{t-1}和 t 时刻上下文向量c_t所决定：

$$s_t=\mathrm{g}(s_{t-1}, y_{t-1}, c_t)$$

其中g为非线性函数。t 时刻上下文向量c_t由源语言隐式状态序列H和注意力模型所产生的权重加权所得：

$$c_t=\sum_{j=1}^{n} a_{t,j} h_j$$

其中注意力模型的权重$a_{t,j}$由 t-1 时刻目标语言隐式状态s_{t-1}与源语言隐式状态序列H产生：

$$a_{t,j}=\frac{\exp(e_{t,j})}{\sum_{k=1}^{n} exp(e_{t,k})}$$

$$e_{t,j}=\mathrm{a}(s_{t-1}, h_j)$$

a为非线性函数。权重$a_{t,j}$可以解释为源语言词语$\mathrm{x_j}$与 t 时刻解码器所产生词语的相关程度。

在取得目标语言隐式状态s_t后，模型通过 softmax 函数估计 t 时刻目标语言单词的概率分布：

$$\mathrm{P}(y_t|y_{<t}, \mathrm{X})=\mathrm{softmax}\left(\mathrm{f}(s_t, y_{t-1}, c_t)\right)$$

接下来，模型可以通过采样或者柱搜索完成 t 时刻词语生成工作，进入 t+1 时刻的工作。解码器一直重复上述解码工作直至生成句子终结符号。

3. 神经机器翻译的译文问题

对比传统统计机器翻译方法，神经机器翻译虽然在翻译性能上更出色，但其译文仍存在如下几个问题：

1）词语表规模受限问题

为了控制模型的时空开销，神经机器翻译通常在源语言端和目标语言端采用规模适当的词语表（词语表规模一般 3 万至 5 万）。对于词语表没有覆盖的词语，模型会将其替换为源语言端和目标语言端相应的 UNK 字符。UNK 字符一方面影响模型完整捕获源语言句子语义信息，另一方面也影响用户理解模型所产生的目标语言句子。这一问题对词形丰富的语言（比如德语）显得尤为严重。（Jean 等[11]）

2）源语言翻译覆盖问题

神经机器翻译在解码过程中通过注意力机制的自动调整，选择关注不同的源语言句子片段来产生对应的目标语言单词。由于缺少约束，注意力机制无法保证源语言句子中的词语被“恰到好处”地关注，从而导致“过翻译”“欠翻译”现象的产生。其中“过翻译”指不该多次翻译的源语言词语被多次翻译，“欠翻译”是指应该被翻译的源语言词语没有被翻译（Tu 等[12]）。

3）翻译不忠实问题

以连续表示的方式进行词语表示，一方面给神经机器翻译带来了更好的泛化能力，另一方面也使得神经机器翻译容易产生不忠实的翻译。此处不忠实的翻译是指模型生成的目标语言词语能够保证目标语言语句的流利度，却无法准确地反映出源语言句子的语义信息（Arthur 等[13]）。

4. 当前研究热点

近年来，神经机器翻译引起了学术界和工业界广泛的关注，在规模受限词语表（Jean 等[11]，Luong 等[14]）、注意力机制（Tu 等[12]，Cohn 等[15]）、神经机器翻译和传统统计机器翻译的结合（He 等[16]，Stahlberg 等[17]）、语言学知识引入（Eriguchi 等[18]，Sennrich 等[19]）、单语语料使用（Gulcehre 等[20]，Sennrich 等[21] ，Cheng 等[22] ，Zhang 等[23] ， Luong 等[24]）、多语言神经机器翻译（Dong 等[25] ，Firat 等[26]）、变分神经机器翻译（Zhang 等[27]）、外部记忆结构（Wang 等[28]）、神经机器翻译模型的训练（Wu 等[8] ，Shen 等[29]）

和模型压缩（See 等[30]）等方面都有相应的研究开展。受限于篇幅，本文只对前三个问题进行讨论。

规模受限词语表问题

受传统统计机器翻译中词语对齐的启发，Luong 等人[14]在目标语言句子中插入定位符号，借助传统统计机器翻译中词对齐信息来定位目标语言中 UNK 符号所对应的源语言单词。在翻译结束后，借助定位信息以查词典的方式处理目标语言句子中的 UNK 符号。Gulcehre 等人[31]观察到翻译译文中很多词语，比如命名实体，是直接从源语言句子中拷贝过来。基于这一现象，他们在神经机器翻译模型上嵌入一种拷贝模式，解码阶段让解码器自动选择是从词语表中选择词语进行生成还是从源语言句子中选择词语进行拷贝。Li 等人[32]提出一种“替换-翻译-恢复”的方法，根据语义相似度对 UNK 字符进行替换，以减轻 UNK 字符对源语言句子结构和语义完整性的破坏。Jean 等人[11]提出一种基于重要性采样的方法，以保证在使用大规模词语表时模型训练复杂度不显著增加。首先分割训练语料为若干个子语料集，在每个子语料集上产生各自单独的小规模词语表。然后在模型训练阶段，对于每个子语料集，利用小规模词语表的计算信息来近似目标函数在大规模词语表上的梯度。

上述神经机器翻译研究都是以词为基本单元进行建模。相比于词单元，细粒度单元的稀疏度要小得多。细粒度单元建模特别是字符单元建模的方法引起了广泛的研究。Costa-jussa 等人[33]以字符表示为底层，采用卷积层和高速层神经网络（convolutional and highway layers）形成词语表示。其实验将所提出的模型布置在源语言，以帮助编码器更完整地捕获源语言句子语义信息，而目标语言仍以单词进行概率估计和生成。Ling 等人[34]进一步在双语两端采用字符表示方法。其以字符表示为底层，采用双向 LSTM 神经网络形成词语表示。解码器中每个词语的生成工作被转化为字符生成工作：解码器生成字符序列，直到生成词语完成符号“EOW”完成单个词语的生成。Sennrich 等人[35]采用字节对码化（Byte Pair Encoding）算法提取子词（subword）单元，完成对单词的拆分。其模型的编码和解码工作均在拆分后的子词上进行。与 Sennrich 等的工作类似，Wu 等人[8]采用算法自动形成的词块（wordpieces）来完成词语拆分工作。Luong 和 Manning[36]提出一种混合字符-词语模型。其中，词语级别模型负责词语表中词

语的建模和生成工作；字符级别模型负责源语言 UNK 符号所对应的单词建模工作和目标语言 UNK 符号的单词恢复工作。Chung 等人[37]跨过词语单元，直接在目标语言进行字符序列生成。其工作以双尺度循环神经网络（Bi-Scale Recurrent Neural Network）作为解码器进行字符解码工作

注意力机制问题

注意力机制极大地提升了神经机器翻译的性能，但是仍存在一些问题。神经机器翻译中注意力机制的完善也是当前的一个研究热点。

一方面，一些研究者让注意力模型在生成当前注意力信息时更充分地考虑注意力的历史信息。Sankaran 等人[38]提出了时序注意力模型。模型在对源语言单词产生注意力权重时，使用该源语言词语自身的历史注意力信息来调整当前的注意力权重。Yang 等人[39]为每个源语言单词配置动态记忆向量并将其引入注意力模型，以帮助注意力模型捕捉注意力历史信息。解码步骤中模型使用每个源语言单词及其周围单词的历史注意力信息更新该单词的动态记忆向量。Wang 等[28]引入外部记忆并通过对外部记忆的读写捕捉目标语言隐式状态历史。解码时，模型从外部记忆中构造出含有更丰富历史信息的中间状态，并使用该中间状态替代目标语言隐式状态完成注意力的生成。Meng 等人[40]设计出一种新颖的交互式注意力机制。在交互式注意力机制中，解码时源端句子向量会被模型依据解码器的状态通过读写进行修改，以帮助模型更好地捕捉源端句子的关注历史。Zhang 等人[41]则绕开注意力机制，通过循环神经网络捕捉上下文向量的历史信息来更好地对源语言句子进行动态关注，以缓解经典注意力模型通过简单加权求和方式得到上下文向量而忽略了词向量内部依赖和非线性关系的问题（见图 4）。

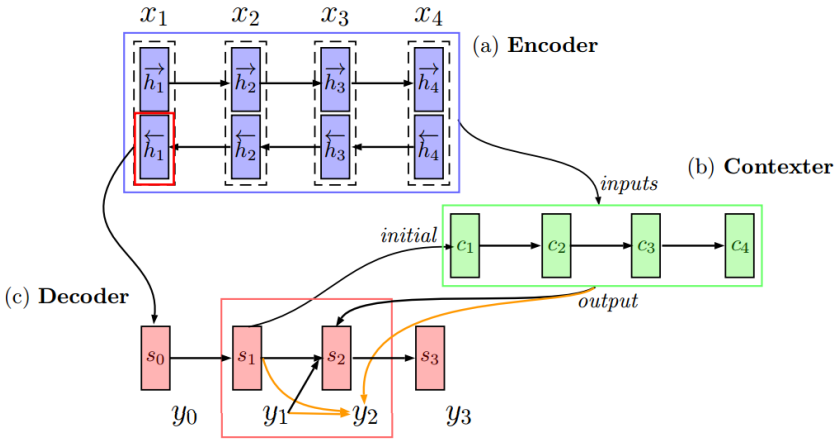

图 4. 循环神经机器翻译结构图[41]

另一方面，注意力模型所产生的注意力可以视作目标语言词语和源语言词语一种对齐信息，一些研究者借鉴传统机器翻译中对齐思想或利用传统机器翻译中的对齐信息对注意力机制进行完善。Cohn[15]等人直接将基于词的统计机器翻译模型中的对齐信息引入注意力模型。具体地，将绝对位置偏差（absolute positional bias）、繁衍度（ fertility）、相对位置偏差（relative position bias）和对齐一致性（alignment consistency）信息作为先验知识直接嵌入注意力模型。Liu 等人[42]和 Mi 等人[43]都将传统机器翻译中词对齐信息作为监督信号引入神经机器翻译训练，利用该监督信号自动地引导注意力模型对注意力进行调整。Feng 等人[44]将传统统计机器翻译的扭曲度概念引入注意力模型。该工作直接将前一时刻的上下文向量信息输入注意力模型，以帮助注意力模型更好地预测目标语言句子的词语顺序。Tu 等人[12]借鉴传统统计机器翻译中的“覆盖（coverage）”概念，在神经机器翻译中引入覆盖模型。该模型为源语言句子每个单词配置一个覆盖向量以存储历史覆盖信息。Cheng 等人[45]则借鉴传统机器翻译中基于一致性对齐的思想，针对基于注意力的神经机器翻译提出了一种基于一致性的训练算法。

神经机器翻译和传统统计机器翻译的结合

传统统计机器翻译有着与神经机器翻译不同的翻译机制，本文第 3 节所述的三个问题在传统统计机器翻译中都不存在，因此传统的统计机器翻译能够对神经机器翻译进行有益的补充。基于这种现象，一些研究者探索将神经机器翻译与传统统计机器翻译进行结合。

He 等人[16]在对数线性模型的框架下对神经机器翻译和传统统计机器翻译进行结合，将神经机器翻译模型和传统统计机器翻译中的翻译模型、单词数惩罚模型和语言模型作为特征置于对数线性模型中。Stahlberg 等人[17]提出一种句法指导的神经机器翻译模型，利用基于层次短语的机器翻译模型来指导神经机器翻译解码。在模型解码时，层次短语模型的翻译假设限制神经机器翻译解码器的搜索空间并调整解码器目标词语预测概率。Arthur 等人[13]通过将估计出的词汇化翻译概率与神经机器翻译的目标语言词语预测概率进行结合，以提高神经机器翻译的译文忠实度。

我们的工作[46]则进一步以神经网络的方式实现神经机器翻译和传统统计机器翻译的深层结合。核心思想是让传统统计机器翻译以类似提词器的方式对神经

机器翻译进行帮助：在解码过程中，首先，根据神经机器翻译的解码信息（目标语言序列片段和注意力历史信息），传统统计机器翻译模型利用统计机器翻译特征进行下一个预测目标词语的精准推荐（图 5 中“SMT model”部分）。然后，基于神经网络的概率估计模型对来自传统统计机器翻译的推荐进行概率估计（图 5 中“$\text{classifier}_{\text{SMT}}$”部分）。最后，基于门机制的概率调节模型使用前一步骤的概率信息对神经机器翻译自身所产生的目标词语预测概率进行相应的调整（图 5 中“gate”部分）。

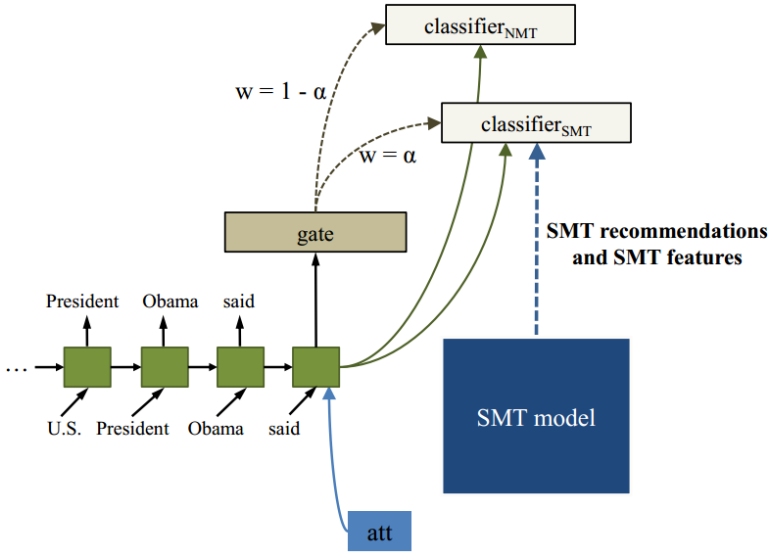

图 5. 融合统计机器翻译推荐的神经机器翻译[46]

5. 总结

神经机器翻译在学术界和工业界的迅猛发展，表明深度学习在机器翻译中的应用取得了成效。神经机器翻译的翻译性能取得了突破，超过了发展多年的传统统计机器翻译，成果振奋人心。然而，我们应该冷静地看到神经机器翻译仍然有很多问题亟待解决。

和其他深度学习所涉及的任务一样，基于深度学习的神经机器翻译较传统统计机器翻译而言，可解释性进一步降低。就模型配置而言，神经机器翻译参数设置和组件配置更多地采用经验性的挑选。在不同语言对、不同词语表规模、不同训练语料规模的情况下，词语表示的维度都经验性地设置为 620 是否合理？如何选择参数配置能保证模型取得最大的功效？这些问题都值得进一步研究讨论。从自然语言处理角度来看，如何从语言学角度提高神经机器翻译的解释性也需要更

多的研究探索。

另外，当前的神经机器翻译完全自动地从双语语料中学习翻译知识，但是对一些外部知识，如双语词典、Wordnet 和知识图谱等，没有过多的触及。Tang 等人[47]和 Zhang 等人[48]都尝试将外部的双语词典引入到神经机器翻译中并取得了初步成效，表明外部知识的引入对神经机器翻译是一个有益的补充。因此，如何有效地将外部知识融入神经机器翻译使其性能进一步提高是一个值得探索的方向。

参考文献

[1] Forcada M L, Ñeco R P. Recursive hetero-associative memories for translation[C]. International Work-Conference on Artificial and Natural Neural Networks: Biological and Artificial Computation: From Neuroscience To Technology. Springer-Verlag, 1997.

[2] Kalchbrenner N, Blunsom P. Recurrent Continuous Translation Models. [C]. EMNLP, 2013.

[3] Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]. EMNLP, 2014.

[4] Sutskever I, Vinyals O, Le Q V, et al. Sequence to Sequence Learning with Neural Networks[C]. NIPS, 2014.

[5] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]. ICLR, 2015.

[6] Junczys-Dowmunt M, Dwojak T, Hoang H. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions[J]. arxiv, 2016.

[7] Bentivogli L, Bisazza A, Cettolo M, et al. Neural versus Phrase-Based Machine Translation Quality: a Case Study[C]. EMNLP, 2016.

[8] Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[J]. arxiv, 2016.

[9] Bojar O, Chatterjee R, Federmann C, et al. Findings of the 2015 Workshop on Statistical Machine Translation[C]. Tenth Workshop on Statistical Machine Translation, 2015.

[10] Cettolo M, Niehues J, Stüker S, et al. The IWSLT 2015 Evaluation Campaign[C]. Twelveth International Workshop, 2015.

[11] Jean S, Cho K, Memisevic R, et al. On Using Very Large Target Vocabulary for Neural Machine Translation[C]. ACL, 2015.

[12] Tu Z, Lu Z, Liu Y, et al. Modeling Coverage for Neural Machine Translation[C]. ACL, 2016.

[13] Arthur P, Neubig G, Nakamura S. Incorporating Discrete Translation Lexicons into Neural Machine Translation[C]. EMNLP, 2016.

[14] Luong M T, Sutskever I, Le Q V, et al. Addressing the Rare Word Problem in Neural Machine Translation[C]. ACL, 2015.

[15] Cohn T, Cong D V H, Vymolova E, et al. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model[C]. NAACL, 2016.

[16] He W, He Z, Wu H, et al. Improved Neural Machine Translation with SMT Features[C]. AAAI,

2016.
[17] Stahlberg F, Hasler E, Waite A, et al. Syntactically Guided Neural Machine Translation[C]. ACL, 2016.
[18] Eriguchi A, Hashimoto K, Tsuruoka Y. Tree-to-Sequence Attentional Neural Machine Translation[C]. ACL, 2016.
[19] Sennrich R, Haddow B. Linguistic Input Features Improve Neural Machine Translation[C]. First Conference on Machine Translation, 2016.
[20] Gulcehre C, Firat O, Xu K, et al. On Using Monolingual Corpora in Neural Machine Translation[J]. Computer Science, 2015.
[21] Sennrich R, Haddow B, Birch A. Improving Neural Machine Translation Models with Monolingual Data[C]. ACL, 2016.
[22] Cheng Y, Xu W, He Z, et al. Semi-Supervised Learning for Neural Machine Translation[C]. ACL, 2016.
[23] Zhang J, Zong C. Exploiting Source-side Monolingual Data in Neural Machine Translation[C]. EMNLP, 2016.
[24] Luong M T, Le Q V, Sutskever I, et al. Multi-task sequence to sequence learning[C]. ICLR, 2016
[25] Dong D, Wu H, He W, et al. Multi-task learning for multiple language translation[C]. ACL, 2015.
[26] Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism[C]. NAACL, 2016.
[27] Zhang B, Xiong D, Su J, et al. Variational Neural Machine Translation[C]. EMNLP, 2016.
[28] Wang M, Lu Z, Li H, et al. Memory-enhanced Decoder for Neural Machine Translation[C]. EMNLP, 2016.
[29] Shen S, Cheng Y, He Z, et al. Minimum Risk Training for Neural Machine Translation[C]. ACL, 2016.
[30] See A, Luong M T, Manning C D. Compression of Neural Machine Translation Models via Pruning[J] . arxiv, 2016.
[31] Gulcehre C, Ahn S, Nallapati R, et al. Pointing the Unknown Words[C]. ACL, 2016.
[32] Li X, Zhang J, Zong C. Towards Zero Unknown Word in Neural Machine Translation[C]. IJCAI, 2016.
[33] Costajussà M R, Fonollosa J A R. Character-based Neural Machine Translation[C]. ACL, 2016.
[34] Ling W, Trancoso I, Dyer C, et al. Character-based Neural Machine Translation[J] . arxiv, 2015.
[35] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units[C]. ACL, 2016.
[36] Luong M T, Manning C D. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models[C]. ACL, 2016.
[37] Chung J, Cho K, Bengio Y. A Character-Level Decoder without Explicit Segmentation for Neural Machine Translation[C]. ACL, 2016.
[38] Sankaran B, Mi H, Al-Onaizan Y, et al. Temporal Attention Model for Neural Machine Translation[J]. arxiv, 2016.
[39] Yang Z, Hu Z, Deng Y, et al. Neural Machine Translation with Recurrent Attention Modeling[J]. arxiv, 2016.
[40] Meng F, Lu Z, Li H, et al. Interactive Attention for Neural Machine Translation[C]. COLING, 2016.
[41] Zhang B, Xiong D, Su J. Recurrent Neural Machine Translation[J]. arxiv, 2016.

[42] Liu L, Utiyama M, Finch A, et al. Neural Machine Translation with Supervised Attention[C]. COLING, 2016.
[43] Mi H, Wang Z, Ittycheriah A. Supervised Attentions for Neural Machine Translation[C]. EMNLP, 2016.
[44] Feng S, Liu S, Li M, et al. Implicit Distortion and Fertility Models for Attention-based Encoder-Decoder NMT Model[J]. arxiv, 2016.
[45] Cheng Y, Shen S, He Z, et al. Agreement-based Joint Training for Bidirectional Attention-based Neural Machine Translation[C]. IJCAI, 2016.
[46] Wang, X, Lu, Z, Tu, Z, et al. Neural Machine Translation Advised by Statistical Machine Translation[J]. arxiv, 2016.
[47] Tang Y, Meng F, Lu Z, et al. Neural Machine Translation with External Phrase Memory [J]. arxiv, 2016.
[48] Zhang J, Zong C. Bridging Neural Machine Translation and Bilingual Dictionaries [J]. arxiv, 2016.