SCIENCE CHINA
Technological Sciences

Special Topic: Natural Language Processing Technology
• Review •

October 2020 Vol. 63 No. 10: 2028–2050
https://doi.org/10.1007/s11431-020-1632-x

Neural machine translation: Challenges, progress and future

ZHANG JiaJun[1,2*] & ZONG ChengQing[1,2,3*]

[1] *National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China;*
[2] *University of Chinese Academy of Sciences, Beijing 100190, China;*
[3] *CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China*

Received March 10, 2020; accepted May 9, 2020; published online September 15, 2020

Machine translation (MT) is a technique that leverages computers to translate human languages automatically. Nowadays, neural machine translation (NMT) which models direct mapping between source and target languages with deep neural networks has achieved a big breakthrough in translation performance and become the de facto paradigm of MT. This article makes a review of NMT framework, discusses the challenges in NMT, introduces some exciting recent progresses and finally looks forward to some potential future research trends.

neural machine translation, Transformer, multimodal translation, low-resource translation, document translation

Citation: Zhang J J, Zong C Q. Neural machine translation: Challenges, progress and future. Sci China Tech Sci, 2020, 63: 2028–2050, https://doi.org/10.1007/s11431-020-1632-x

1 Introduction

The concept of machine translation (MT) was formally proposed in 1949 by Weaver [1] who believed it is possible to use modern computers to automatically translate human languages. From then on, machine translation has become one of the most challenging task in the area of natural language processing and artificial intelligence. Many researchers of several generations dedicated themselves to realize the dream of machine translation.

From the viewpoint of methodology, approaches to MT mainly fall into two categories: rule-based method and data-driven approach. Rule-based methods were dominant and preferable before 2000s. In this kind of methods, bilingual linguistic experts are responsible to design specific rules for source language analysis, source-to-target language transformation and target language generation. Since it is very subjective and labor intensive, rule-based systems are difficult to be scalable and they are fragile when rules cannot cover the unseen language phenomena.

In contrast, the data-driven approach aims at teaching computers to learn how to translate from lots of human-translated parallel sentence pairs (parallel corpus). The study of data-driven approach has experienced three periods. In the middle of 1980s, ref. [2] proposed example-based MT which translates a sentence by retrieving the similar examples in the human-translated sentence pairs.

From early 1990s, statistical machine translation (SMT) has been proposed and word or phrase level translation rules can be automatically learned from parallel corpora using probabilistic models [3–5]. Thanks to the availability of more and more parallel corpora, sophisticated probabilistic models such as noisy channel model and log-linear model achieve better and better translation performance. Many companies (e.g., Google, Microsoft and Baidu) have developed online SMT systems which much benefit the users. However, due to complicated integration of multiple manually designed components such as translation model, language model and reordering model, SMT cannot make full use of large-scale parallel corpora and translation quality is far from satisfactory.

*Corresponding authors (email: jjzhang@nlpr.ia.ac.cn; cqzong@nlpr.ia.ac.cn)

No breakthrough has been achieved more than 10 years until the introduction of deep learning into MT. Since 2014, neural machine translation (NMT) based on deep neural networks has quickly developed [6–9]. In 2016, through extensive experiments on various language pairs, refs. [10, 11] demonstrated that NMT has made a big breakthrough and obtained remarkable improvements compared to SMT, and even approached human-level translation quality [12]. This article attempts to give a review of NMT framework, discusses some challenging research tasks in NMT, introduces some exciting progresses and forecasts several future research topics.

The remainder of this article is organized as follows. Sect. 2 first introduces the background and state-of-the-art paradigm of NMT. In Sect. 3 we discuss the key challenging research tasks in NMT. From Sect. 4 to Sect. 7, the recent progresses are presented concerning each challenge. Sect. 8 discusses the current state of NMT compared to expert translators and finally looks forward to some potential research trends in the future.

2 Neural machine translation

2.1 Encoder-decoder framework

Neural machine translation is an end-to-end model following an encoder-decoder framework that usually includes two neural networks for encoder and decoder respectively [6–9]. As shown in Figure 1, the encoder network first maps each input token of the source-language sentence into a low-dimensional real-valued vector (aka word embedding) and then encodes the sequence of vectors into distributed semantic representations, from which the decoder network generates the target-language sentence token by token [1] from left to right.

From the probabilistic perspective, NMT models the conditional probability of the target-language sentence $\boldsymbol{y} = y_0, \cdots, y_i, \cdots, y_I$ given the source-language sentence $\boldsymbol{x} = x_0, \cdots, x_j, \cdots, x_J$ as a product of token-level translation probabilities.

$$P(\boldsymbol{y}|\boldsymbol{x}, \theta) = \prod_{i=0}^{I} P(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i}, \theta), \tag{1}$$

where $\boldsymbol{y}_{<i} = y_0, \cdots, y_{i-1}$ is the partial translation which has been generated so far. x_0, y_0 and x_J, y_I are often special symbols <s> and </s> indicating the start and end of a sentence respectively.

The token-level translation probability can be defined as follows:

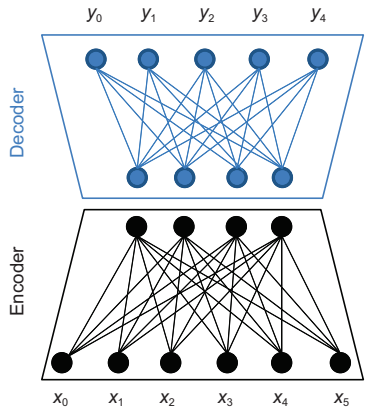

Figure 1 (Color online) Encoder-decoder framework for neural machine translation. The encoder encodes the input sequence $x_0x_1x_2x_3x_4x_5$ into distributed semantic representations based on which the decoder produces an output sequence $y_0y_1y_2y_3y_4$.

$$P(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i}, \theta) = \frac{\exp\Big(g(\boldsymbol{x}, \boldsymbol{y}_{<i}, y_i, \theta)\Big)}{\sum_{y' \in V} \exp\Big(g(\boldsymbol{x}, \boldsymbol{y}_{<i}, y', \theta)\Big)}, \tag{2}$$

in which V denotes the vocabulary of the target language and $g(\cdot)$ is a non-linear function that calculates a real-valued score for the prediction y_i conditioned on the input $\boldsymbol{x}$, the partial translation $\boldsymbol{y}_{<i}$ and the model parameters θ. The non-linear function $g(\cdot)$ is realized through the encoder and decoder networks. The input sentence $\boldsymbol{x}$ is abstracted into hidden semantic representations $\boldsymbol{h}$ through multiple encoder layers. $\boldsymbol{y}_{<i}$ is summarized into the target-side history context representation $\boldsymbol{z}$ with decoder network which further combines $\boldsymbol{h}$ and $\boldsymbol{z}$ using an attention mechanism to predict the score of y_i.

The network parameters θ can be optimized to maximize the log-likelihood over the bilingual training data $D = \{(\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)})\}_{m=1}^{M}$:

$$\theta = \text{argmax}_{\theta^*} \sum_{m=1}^{M} \log P(\boldsymbol{y}^{(m)}|\boldsymbol{x}^{(m)}, \theta^*). \tag{3}$$

These years have witnessed the fast development of the encoder-decoder networks from recurrent neural network [6, 7], to convolutional neural network [8] and then to self-attention based neural network Transformer [9]. At present, Transformer is the state-of-the-art in terms of both quality and efficiency.

2.2 Transformer

In Transformer[2], the encoder includes N identical layers and each layer is composed of two sub-layers: the self-attention sub-layer followed by the feed-forward sub-layer, as shown

1) Currently, subword is the most popular translation token for NMT [13].
2) Model and codes can be found at https://github.com/tensorflow/tensor2tensor.

in the left part of Figure 2. The self-attention sub-layer calculates the output representation of a token by attending to all the neighbors in the same layer, computing the correlation score between this token and all the neighbors, and finally linearly combining all the representations of the neighbors and itself. The output of the N-th encoder layer is the source-side semantic representation $\boldsymbol{h}$. The decoder as shown in the right part in Figure 2 also consists of N identical layers. Each layer has three sub-layers. The first one is the masked self-attention mechanism that summarizes the partial prediction history. The second one is the encoder-decoder attention sub-layer determining the dynamic source-side contexts for current prediction and the third one is the feed-forward sub-layer. Residual connection and layer normalization are performed for each sub-layer in both of the encoder and decoder.

It is easy to notice that the attention mechanism is the key component. There are three kinds of attention mechanisms, including encoder self-attention, decoder masked self-attention and encoder-decoder attention. They can be formalized into the same formula.

$$\text{Attention}(\boldsymbol{q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{4}$$

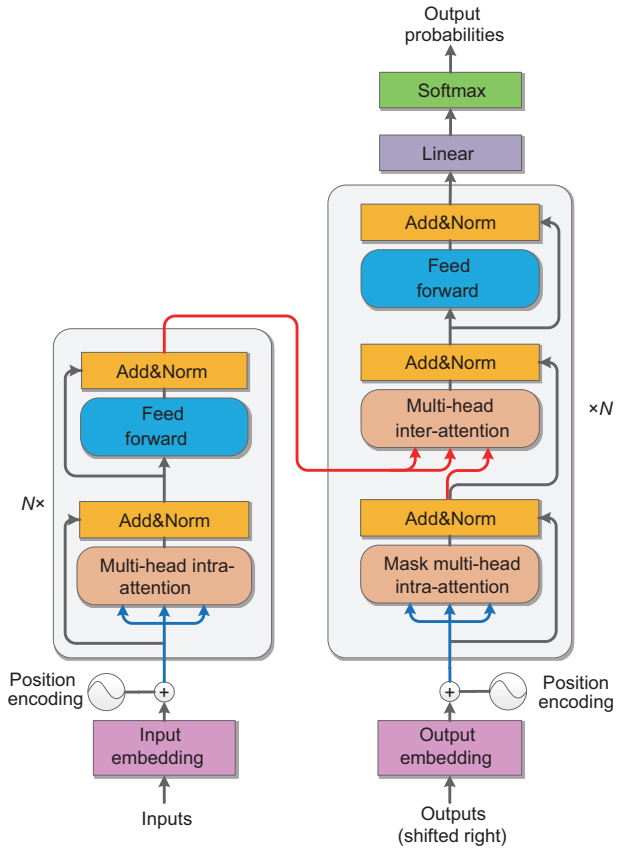

Figure 2 (Color online) The Transformer architecture in which attention mechanism is the core in both of the encoder and decoder networks. *shift right* means that the prediction of the previous time-step will shift right as the input context to predict next output token.

where $\boldsymbol{q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ stand for a query, the key list and the value list respectively. d_k is the dimension of the key.

For the encoder self-attention, the queries, keys and values are from the same layer. For example, considering we calculate the output of the first layer in the encoder at the j-th position, let $\boldsymbol{x}_j$ be the sum vector of input token embedding and the positional embedding. The query is vector $\boldsymbol{x}_j$. The keys and values are the same and both are the embedding matrix $\boldsymbol{x} = [\boldsymbol{x}_0 \cdots \boldsymbol{x}_J]$. Then, multi-head attention is proposed to calculate attentions in different subspaces.

$$\begin{aligned}\text{MultiHead}(q,\boldsymbol{K},\boldsymbol{V}) &= \text{Concat}_i(head_i)\boldsymbol{W}_O,\\ \text{head}_i &= \text{Attention}(\boldsymbol{q}\boldsymbol{W}_Q^i, \boldsymbol{K}\boldsymbol{W}_K^i, \boldsymbol{V}\boldsymbol{W}_V^i),\end{aligned} \tag{5}$$

in which $\boldsymbol{W}_Q^i$,, $\boldsymbol{W}_K^i$, $\boldsymbol{W}_V^i$ and $\boldsymbol{W}_O$ denote projection parameter matrices .

Using eq. (5) followed by residential connection, layer normalization and a feed-forward network, we can get the representation of the second layer. After N layers, we obtain the input contexts $\boldsymbol{C} = [\boldsymbol{h}_0, \cdots, \boldsymbol{h}_J]$.

The decoder masked self-attention is similar to that of encoder except that the query at the i-th position can only attend to positions before i, since the predictions after i-th position are not available in the auto-regressive left-to-right unidirectional inference.

$$z_i = \text{Attention}(\boldsymbol{q}_i, \boldsymbol{K}_{\leq i}, \boldsymbol{V}_{\leq i}) = \text{softmax}\left(\frac{\boldsymbol{q}_i\boldsymbol{K}_{\leq i}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_{\leq i}. \tag{6}$$

The encoder-decoder attention mechanism is to calculate the source-side dynamic context which is responsible to predict the current target-language token. The query is the output of the masked self-attention sub-layer z_i. The keys and values are the same encoder contexts $\boldsymbol{C}$. The residential connection, layer normalization and feed-forward sub-layer are then applied to yield the output of a whole layer. After N such layers, we obtain the final hidden state z_i. Softmax function is then employed to predict the output y_i, as shown in the upper right part of Figure 2.

3 Key challenging research tasks

Although Transformer has significantly advanced the development of neural machine translation, many challenges still remain to be addressed. Obviously, designing better NMT framework must be the most important challenge. However, since the innovation of Transformer, almost no more effective NMT architecture has been proposed. Ref. [14] presented an alternative encoder-decoder framework RNMT+ which combines the merits of RNN-based and Transformer-based models to perform translation. Refs. [15, 16] investigated how to design much deeper Transformer model and ref. [17] presented a Reformer model enabling rich interaction between

encoder and decoder. Ref. [18] attempted to replace self-attention with dynamic convolutions. Ref. [19] proposed the evolved Transformer using neural architecture search. Ref. [20] aimed to improve Transformer from the perspective of multi-particle dynamic system. Note that these models do not introduce big change on the NMT architecture. Pursuing to design novel and more effective NMT framework will be a long way to go. In this section, we analyze and discuss the key challenges[3)] facing NMT from its formulation.

From the introduction in Sect. 2.1, NMT is formally defined as a sequence-to-sequence prediction task in which four assumptions are hidden in default. First, the input is a sentence rather than paragraphs and documents. Second, the output sequence is generated in a left-to-right autoregressive manner. Third, the NMT model is optimized over the bilingual training data which should include large-scale parallel sentences in order to learn good network parameters. Fourth, the processing objects of NMT are the pure texts (tokens, words and sentences) instead of speech and videos. Accordingly, four key challenges can be summarized as follows.

(1) Document neural machine translation. In NMT formulation, sentence is the basic input for modeling. However, some words in the sentence are ambiguous and the sense can only be disambiguated with the context of surrounding sentences or paragraphs. And when translating a document, we need to guarantee the same terms in different sentences lead to the same translation while performing translation sentence by sentence independently cannot achieve this goal. Moreover, many discourse phenomena such as coreference, omissions and coherence, cannot be handled in the absence of document-level information. Obviously, it is a big challenge how to take full advantage of contexts beyond sentences in neural machine translation.

(2) Non-autoregressive decoding and bidirectional inference. Left-to-right decoding token by token follows an autoregressive style which seems natural and is in line with human reading and writing. It is also easy for training and inference. However, it has several drawbacks. On one hand, the decoding efficiency is quite limited since the i-th translation token can be predicted only after all the previous $i-1$ predictions have been generated. On the other hand, predicting the i-th token can only access the previous history predictions while cannot utilize the future context information in autoregressive manner, leading to inferior translation quality. Thus, it is a challenge how to break the autoregressive inference constraint. Non-autoregressive decoding and bidirectional inference are two solutions from the perspectives of efficiency and quality respectively.

(3) Low-resource translation. There are thousands of human languages in the world and abundant bitexts are only available in a handful of language pairs such as English-German, English-French and English-Chinese. Even in the resource-rich language pair, the parallel data are unbalanced since most of the bitexts mainly exist in several domains (e.g., news and patents). That is to say, the lack of parallel training corpus is very common in most languages and domains. It is well-known that neural network parameters can be well optimized on highly repeated events (frequent word/phrase translation pairs in the training data for NMT) and the standard NMT model will be poorly learned on low-resource language pairs. As a result, how to make full use of the parallel data in other languages (pivot-based translation and multilingual translation) and how to take full advantage of non-parallel data (semi-supervised translation and unsupervised translation) are two challenges facing NMT.

(4) Multimodal neural machine translation. Intuitively, human language is not only about texts and understanding the meaning of a language may need the help of other modalities such as speech, image and videos. Concerning the well-known example that determines the meaning of the word bank when translating the sentence "he went to the bank", it will be correctly translated if we are given an image in which a man is approaching a river. Furthermore, in many scenarios, we are required to translate a speech or a video. For example, simultaneous speech translation is more and more demanding in various conferences or international live events. Therefore, how to perform multimodal translation under the encoder-decoder architecture is a big challenge of NMT. How to make full use of different modalities in multimodal translation and how to balance the quality and latency in simultaneous speech translation are two specific challenges.

In the following sections, we briefly introduce the recent progress for each challenge.

4 Document-level neural machine translation

As we discussed in Sect. 3 that performing translation sentence by sentence independently would introduce several risks. An ambiguous word may not be correctly translated without the necessary information in the surrounding contextual sentences. A same term in different sentences in the same document may result in inconsistent translations. Furthermore, many discourse phenomena, such as coreference, omissions and cross-sentence relations, cannot be well handled. In a word, sentence-level translation will harm the coherence and cohesion of the translated documents if we ignore the discourse connections and relations between sentences.

3) Refs. [21–23] have also discussed various challenges.

In general, document-level machine translation (docMT) aims at exploiting the useful document-level information (multiple sentences around the current sentence or the whole document) to improve the translation quality of the current sentence as well as the coherence and cohesion of the translated document. docMT has already been extensively studied in the era of statistical machine translation (SMT), in which most researchers mainly propose explicit models to address some specific discourse phenomena, such as lexical cohesion and consistency [26–28], coherence [29] and coreference [30]. Due to complicate integration of multiple components in SMT, these methods modeling discourse phenomenon do not lead to promising improvements.

The NMT model dealing with semantics and translation in the distributed vector space facilitates the use of wider and deep document-level information under the encoder-decoder framework. It does not need to explicitly model specific discourse phenomenon as that in SMT. According to the types of used document information, document-level neural machine translation (docNMT) can roughly fall into three categories: dynamic translation memory [31, 32], surrounding sentences [24, 25, 33–37] and the whole document [38–40].

Ref. [32] presented a dynamic cache-like memory to maintain the hidden representations of previously translated words. The memory contains a fixed number of cells and each cell is a triple $(\boldsymbol{c}_t, \boldsymbol{s}_t, y_t)$ where y_t is the prediction at the t-th step, $\boldsymbol{c}_t$ is the source-side context representation calculated by the attention model and $\boldsymbol{s}_t$ is the corresponding decoder state. During inference, when predicting the i-th prediction for a test sentence, $\boldsymbol{c}_i$ is first obtained through attention model and the probability $p(\boldsymbol{c}_t|\boldsymbol{c}_i)$ is computed based on their similarity. Then memory context representation $\boldsymbol{m}_i$ is calculated by linearly combining all the values $\boldsymbol{s}_t$ with $p(\boldsymbol{c}_t|\boldsymbol{c}_i)$. This cache-like memory can encourage the words in similar contexts to share similar translations so that cohesion can be enhanced to some extent.

The biggest difference between the use of whole document and surrounding sentences lies in the number of sentences employed as the context. This article mainly introduces the methods exploiting surrounding sentences for docNMT. Relevant experiments further show that subsequent sentences on the right contribute little to the translation quality of the current sentence. Thus, most of the recent work aim at fully exploring the previous sentences to enhance docNMT. These methods can be divided into two categories. One just utilizes the previous source-side sentences [24, 33, 34, 41]. The other uses the previous source sentences as well as their target translations [25, 36].

If only previous source-side sentences are leveraged, the previous sentences can be concatenated with the current sentence as the input to the NMT model [41] or could be encoded into a summarized source-side context with a hierarchical neural network [34]. Ref. [24] presented a cascaded attention model to make full use of the previous source sentences. As shown in Figure 3(a) [24, 25], previous sentence is first encoded as the document-level context representation. When encoding the current sentence, each word will attend to the document-level context and obtain a context-enhanced source representation. During the calculation of cross-language attention in the decoder, the current source sentences together with the document-level context are both leveraged to predict the target word. The probability of translation sentence

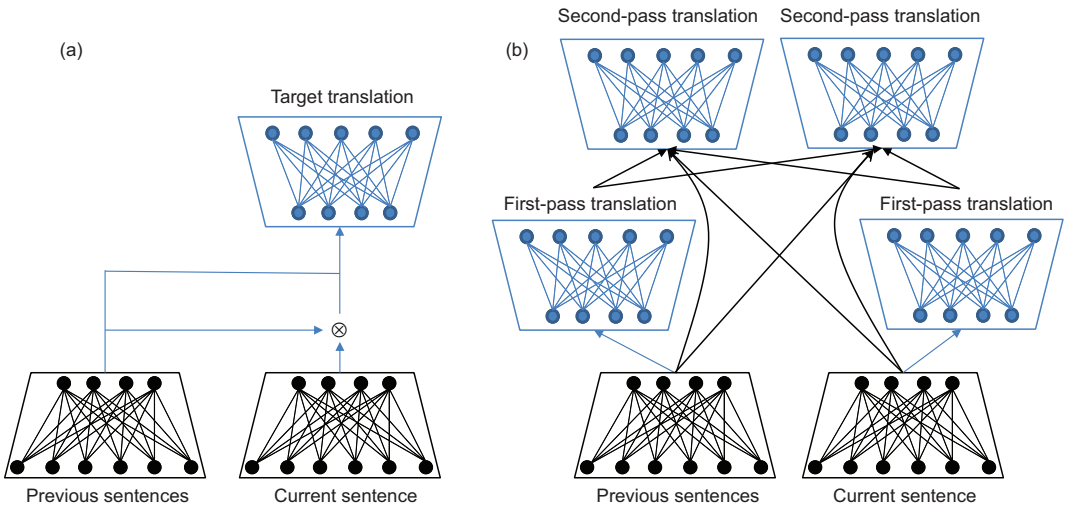

Figure 3 (Color online) Illustration of two docNMT models. The left part shows the cascaded attention model proposed by ref. [24] in which the previous source sentences are first leveraged to enhance the representation of current source sentence and then used again in the decoder. The right part illustrates the two-pass docNMT model proposed by ref. [25] in which sentence-level NMT first generates preliminary translation for each sentence and then the first-pass translations together with the source-side sentences are employed to generate the final translation results.

given the current sentence and the previous context sentences is formulated as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}, \mathrm{doc}_x; \theta) = \prod_{i=0}^{I} P(y_i|\boldsymbol{y}_{<i}, \boldsymbol{x}, \mathrm{doc}_x; \theta), \tag{7}$$

where doc_x denotes the source-side document-level context, namely previous sentences.

If both of previous source sentences and their translations are employed, two-pass decoding is more suitable for the docNMT model [25]. As illustrated in Figure 3(b), the sentence-level NMT model can generate preliminary translations for each sentence in the first-pass decoding. Then, the second-pass model will produce final translations with the help of source sentences and their preliminary translation results. The probability of the target sentence in the second pass can be written by

$$P(\boldsymbol{y}|\boldsymbol{x}, \mathrm{doc}_x, \mathrm{doc}_y; \theta) = \prod_{i=0}^{I} P(y_i|\boldsymbol{y}_{<i}, \boldsymbol{x}, \mathrm{doc}_x, \mathrm{doc}_y; \theta), \tag{8}$$

in which doc_y denotes the first-pass translations of doc_x.

Since most methods for docNMT are designed to boost the overall translation quality (e.g., BLEU score), it still remains a big problem whether these methods indeed well handle the discourse phenomena. To address this issue, ref. [42] conducted an empirical investigation of the docNMT model on the performance of processing various discourse phenomena, such as coreference, cohesion and coherence. Their findings indicate that multi-encoder model exploring only the source-side previous sentences performs poorly in handling the discourse phenomena while exploiting both source sentences and target translations leads to the best performance. Accordingly, ref. [43, 44] recently focused on designing better document-level NMT to improve on specific discourse phenomena such as deixis, ellipsis and lexical cohesion for English-Russian translation.

5 Non-autoregressive decoding and bidirectional inference

Most NMT models follow the autoregressive generation style which produces output word by word from left to right. Just as Sect. 3 discussed, this paradigm has to wait for $i-1$ time steps before starting to predict the i-th target word. Furthermore, left-to-right autoregressive decoding cannot exploit the target-side future context (future predictions after i-th word). Recently, many research work attempt to break this decoding paradigm. Non-autoregressive Transformer (NAT) [45] is proposed to remarkably lower down the latency by emitting all of the target words at the same time and bidirectional inference [46, 47] is introduced to improve the translation quality by making full use of both history and future contexts.

5.1 Non-autoregressive decoding

Non-autoregressive Transformer (NAT) aims at producing an entire target output in parallel. Different from the autoregressive Transformer model (AT) which terminates decoding when emitting an end-of-sentence token $\langle /s \rangle$, NAT has to know how many target words should be generated before parallel decoding. Accordingly, NAT calculates the conditional probability of a translation $\boldsymbol{y}$ given the source sentence $\boldsymbol{x}$ as follows:

$$P_{\mathrm{NAT}}(\boldsymbol{y}|\boldsymbol{x}; \theta) = P_L(I|\boldsymbol{x}; \theta) \cdot \prod_{i=0}^{I} P(y_i|\boldsymbol{x}; \theta). \tag{9}$$

To determine the output length, ref. [45] proposed to use the fertility model which predicts the number of target words that should be translated for each source word. We can perform word alignment on the bilingual training data to obtain the gold fertilities for each sentence pair. Then, the fertility model can be trained together with the translation model. For each source word x_j, suppose the predicted fertility is $\Phi(x_j)$. The output length will be $I = \sum_{j=0}^{J} \Phi(x_j)$.

Another issue remains that AT let the previous generated output y_{i-1} be the input at the next time step to predict the i-th target word but NAT has no such input in the decoder network. Ref. [45] found that translation quality is particularly poor if omitting the decoder input in NAT. To address this, they resort to the fertility model again and copy each source word as many times as its fertility $\Phi(x_j)$ into the decoder input. The empirical experiments show that NAT can dramatically boost the decoding efficiency by 15× speedup compared to AT. However, NAT severely suffers from accuracy degradation.

The low translation quality may be due to at least two critical issues of NAT. First, there is no dependency between target words although word dependency is ubiquitous in natural language generation. Second, the decoder inputs are the copied source words which lie in different semantic space with target words. Recently, to address the shortcomings of the original NAT model, several methods are proposed to improve the translation quality of NAT while maintaining its efficiency [48–53].

Ref. [49] proposed a semi-autoregressive Transformer model (SAT) to combine the merits of both AT and NAT. SAT keeps the autoregressive property in global but performs NAT in local. Just as shown in Figure 4, SAT generates K successive target words at each time step in parallel. If $K = 1$, SAT will be exactly AT. It will become NAT if $K = I$. By choosing an appropriate K, dependency relation between fragments is well modeled and the translation quality can be much improved with some loss of efficiency.

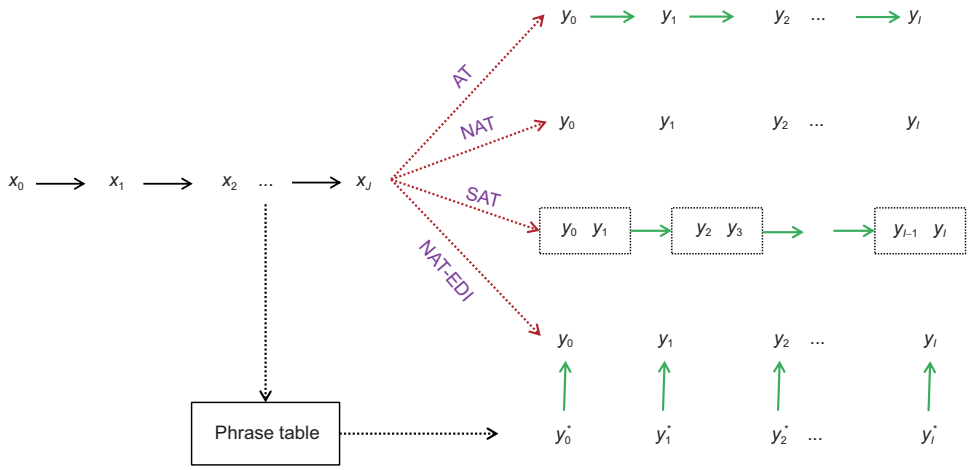

Figure 4 (Color online) Illustration of autoregressive NMT model and various non-autoregressive NMT models. **AT** denotes the conventional autoregressive NMT paradigm in which the i-th prediction can fully utilize the partial translation of $i-1$ words. **NAT** indicates the non-autoregreesive NMT model that generates all the target words simultaneously. **SAT** is a variant of **NAT** which produces an ngram each time. **NAT-EDI** denotes the non-autoregressive NMT model with enhanced decoder input which is generated by retrieving the phrase table.

To mimic the decoder input in the AT model, ref. [50] introduced a simple but effective method that employs a phrase table which is the core component in SMT to convert source words into target words. Specifically, they first greedily segment the source sentence into phrases with maximum match algorithm. Suppose the longest phrase in the phrase table contains K words. $\boldsymbol{x}_{0:K-1}$ is a phrase if it matches an entry in the phrase table. Otherwise they iteratively check $\boldsymbol{x}_{0:K-2}$, $\boldsymbol{x}_{0:K-3}$ and so on. If $\boldsymbol{x}_{0:h}$ is a phrase, then they start to check $\boldsymbol{x}_{h+1:h+K}$. After segmentation, each source phrase is mapped into target translations which are concatenated together as the new decoder input, as shown in Figure 4. Due to proper modeling of the decoder input with a highly efficient strategy, translation quality is substantially improved while the decoding speed is even faster than baseline NAT.

5.2 Bidirectional inference

From the viewpoint of improving translation quality, autoregressive model can be enhanced by exploring the future context on the right. In addition to predicting and estimating the future contexts with various models [54–56], researchers find that left-to-right (L2R) and right-to-left (R2L) autoregressive models can generate complementary translations [46, 47, 57, 58]. For example, in Chinese-to-English translation, experiments show that L2R can generate better prefix while R2L is good at producing suffix. Intuitively, it is a promising direction to combine the merits of bidirectional inferences and fully exploit both history and future contexts on the target side.

To this end, many researchers resort to exploring bidirectional decoding to take advantages of both L2R and R2L inferences. These methods are mainly fall into four categories: (1) enhancing agreement between L2R and R2L predictions [57, 59]; (2) reranking with bidirectional decoding [57, 60, 61]; (3) asynchronous bidirectional decoding [46, 62] and (4) synchronous bidirectional decoding [47, 63, 64].

Ideally, L2R decoding should generate the same translation as R2L decoding. Under this reasonable assumption, refs. [57, 59] introduced agreement constraint or regularization between L2R and R2L predictions during training. Then, L2R inference can be improved.

The reranking algorithm is widely used in machine translation, and the R2L model can provide an estimation score for the quality of L2R translation from another parameter space [57, 60, 61]. Specifically, L2R first generates a n-best list of translations. The R2L model is then leveraged to force decode each translation leading to a new score. Finally, the best translation is selected according to the new scores.

Refs. [46, 62] proposed an asynchronous bidirectional decoding model (ASBD) which first obtains the R2L outputs and optimizes the L2R inference model based on both of the source input and the R2L outputs. Specifically, ref. [46] first trained a R2L model with the bilingual training data. Then, the optimized R2L decoder translates the source input of each sentence pair and produces the outputs (hidden states) which serve as the additional context for L2R prediction when optimizing the L2R inference model. Due to explicit use of right-side future contexts, the ASBD model significantly improves the translation quality. But these approaches still suffer from two issues. On one hand, they have to train two separate NMT models for L2R and R2L inferences respectively. And the two-pass decoding strategy makes the latency much increased. On the other hand, the two models cannot interact

with each other during inference, which limits the potential of performance improvement.

Ref. [47] proposed a synchronous bidirectional decoding model (SBD) that produces outputs using both L2R and R2L decoding simultaneously and interactively. Specifically, a new synchronous attention model is proposed to conduct interaction between L2R and R2L inferences. The top part in Figure 5 gives a simple illustration of the proposed synchronous bidirectional inference model. The dotted arrows on the target side is the core of the SBD model. L2R and R2L inferences interact with each other in an implicit way illustrated by the dotted arrows. All the arrows indicate the information passing flow. Solid arrows show the conventional history context dependence while dotted arrows introduce the future context dependence on the other inference direction. For example, besides the past predictions (y_0^{l2r}, y_1^{l2r}), L2R inference can also utilize the future contexts (y_0^{r2l}, y_1^{r2l}) generated by the R2L inference when predicting y_2^{l2r}. The conditional probability of the translation can be written as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \begin{cases} \prod_{i=0}^{I} P(\overrightarrow{y_i}|\overrightarrow{y}_0 \cdots \overrightarrow{y}_{i-1}, x, \overleftarrow{y}_0 \cdots \overleftarrow{y}_{i-1}), & \text{if L2R,} \\ \prod_{i=0}^{I'-1} P(\overleftarrow{y_i}|\overleftarrow{y}_0 \cdots \overleftarrow{y}_{i-1}, x, \overrightarrow{y}_0 \cdots \overrightarrow{y}_{i-1}), & \text{if R2L.} \end{cases} \tag{10}$$

To accommodate L2R and R2L inferences at the same time, they introduced a novel beam search algorithm. As shown in the bottom right of Figure 5, at each timestep during decoding, each half beam maintains the hypotheses from L2R and R2L decoding respectively and each hypothesis is generated by leveraging already predicted outputs from both directions. At last, the final translation is chosen from L2R and R2L results according to their translation probability. Thanks to appropriate rich interaction, the SBD model substantially boosts the translation quality while the decoding speed is only 10% slowed down.

Ref. [63] further noticed that L2R and R2L are not necessary to produce the entire translation sentence. They let L2R generate the left half translation and make R2L produce the right half, and then two halves are concatenated to form the final translation. Using proper training algorithms, they demonstrated through extensive experiments that both translation quality and decoding efficiency can be significantly improved compared to the baseline Transformer model.

6 Low-resource translation

Most NMT models assume that enough bilingual training data are available, which is the rare case in real life. For a low-resource language pair, a natural question may arise that what kind of knowledge can be transferred to build a relatively good NMT system. This section will discuss three kinds of methods. One attempts to share translation knowledge from other resource-rich language pairs, in which pivot translation and multilingual translation are the two key techniques. Pivot translation assumes that for the low-resource pair A and B, there is a language C that has rich bitexts with A and B, respectively [65, 66]. This section mainly discusses the technique of multilingual translation in the first category. The second kind of methods resort to semi-supervised approach which takes full advantages of limited bilingual training data and abundant monolingual data. The last one leverages unsupervised algorithm that requires monolingual data

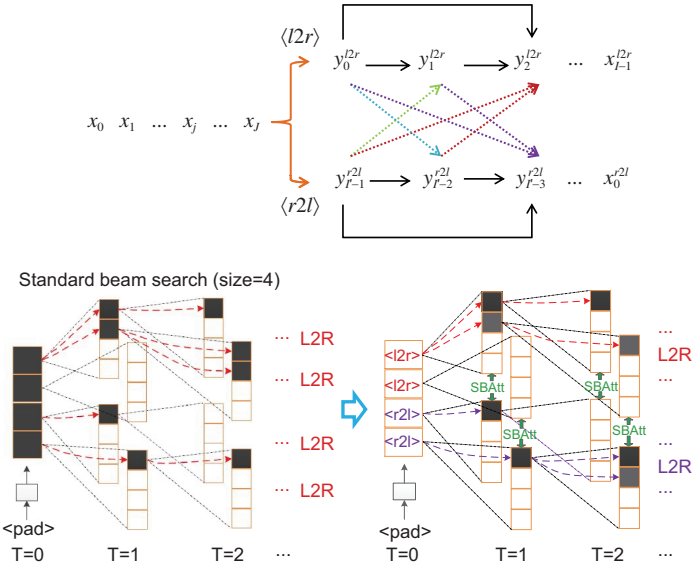

Figure 5 (Color online) Illustration of the synchronous bidirectional inference model. The top demonstrates how the bidirectional contexts can be leveraged during inference. The bottom compares the beam search algorithm between the conventional NMT and the synchronous bidirectional NMT.

only.

6.1 Multilingual neural machine translation

Let us first have a quick recap about the NMT model based on encoder-decoder framework. The encoder is responsible for mapping the source language sentence into distributed semantic representations. The decoder is to convert the source-side distributed semantic representations into target language sentence. Apparently, the encoder and the decoder (excluding the cross-language attention component) are just single-language dependent. Intuitively, the same source language in different translation systems (e.g., Chinese-to-English, Chinese-to-Hindi) can share the same encoder and the same target language can share the same decoder (e.g., Chinese-to-English and Hindi-to-English). Multilingual neural machine translation is a framework that aims at building a unified NMT model capable of translating multiple languages through parameter sharing and knowledge transferring.

Ref. [67] is the first to design a multi-task learning method which shares the same encoder for one-to-many translation (one source language to multiple target languages). Ref. [68] proposed to share the decoder for many-to-one translation (many source languages to one target language). Refs. [69, 70] proposed to share attention mechanism for many-to-many translation (many source languages to many target languages). Despite performance improved for low-resource languages, these methods are required to design a specific encoder or decoder for each language, hinders their scalability in dealing with many languages.

Ref. [71] goes a step further and let all source languages share the same encoder and all the target languages share the same decoder. They have successfully trained a single encoder-decoder NMT model for multilingual translation. The biggest issue is that the decoder is unaware of which target language should be translated to at the test phase. To this end, ref. [71] introduced a simple strategy and added a special token indicating target language (e.g, 2en and 2zh) at the beginning of the source sentence. By doing this, low-resource languages have the biggest chance to share translation knowledge from other resource-rich languages. It also enables zero-shot translation as long as the two languages are employed as source and target in the multilingual NMT model. In addition, this unified multilingual NMT is very scalable and could translate all the languages in one model ideally. However, experiments find that the output is sometimes mixed of multiple languages even using a translation direction indicator. Furthermore, this paradigm enforces different source/target languages to share the same semantic space, without considering the structural divergency among different languages. The consequence is that the single model based multilingual NMT yields inferior translation performance compared to individually trained bilingual counterparts. Most of recent research work mainly focus on designing better models to well balance the language-independent parameter sharing and the language-sensitive module design.

Ref. [72] augmented the attention mechanism in decoder with language-specific signals. Ref. [73] proposed to use language-sensitive positions and language-dependent hidden presentations for one-to-many translation. Ref. [74] designed an algorithm to generate language-specific parameters. Ref. [75] designed a language clustering method and forced languages in the same cluster to share the parameters in the same semantic space. Ref. [76] attempted to generate two languages simultaneously and interactively by sharing encoder parameters. Ref. [77] proposed a compact and language-sensitive multilingual translation model which attempts to share most of the parameters while maintaining the language discrimination.

As shown in Figure 6, ref. [77] designed four novel modules in the Transformer framework compared with single-model based multilingual NMT. First, they introduced a representor to replace both encoder and decoder by sharing weight parameters of the self-attention block, feed-forward and normalization blocks (middle part in Figure 6). It makes the multilingual NMT model as compact as possible and maximizes the knowledge sharing among different languages. The objective function over L language pairs becomes

$$\mathcal{L}(\theta) = \sum_{l=1}^{L} \sum_{m=1}^{M_l} \log P(\boldsymbol{y}_l^{(m)} | \boldsymbol{x}_l^{(m)}; \theta_{\text{rep}}, \theta_{\text{att}}), \tag{11}$$

where θ_{rep} and θ_{att} denote parameters of representor and attention mechanism respectively.

However, the representor further reduces the ability to discriminate different languages. To address this, they introduced three language-sensitive modules.

(1) Language-sensitive embedding (bottom part in Figure 6): they compared four categories of embedding sharing patterns, namely language-based pattern (different languages have separate input embeddings), direction-based patter (languages in source side and target side have different input embeddings), representor-based pattern (shared input embeddings for all languages) and three-way weight tying pattern proposed by ref. [78], in which the output embedding of the target side is also shared besides representor-based sharing.

(2) Language-sensitive attention (middle part in Figure 6): this mechanism allows the model to select the cross-lingual attention parameters according to specific translation tasks dynamically.

(3) Language-sensitive discriminator (top part in Figure 6):

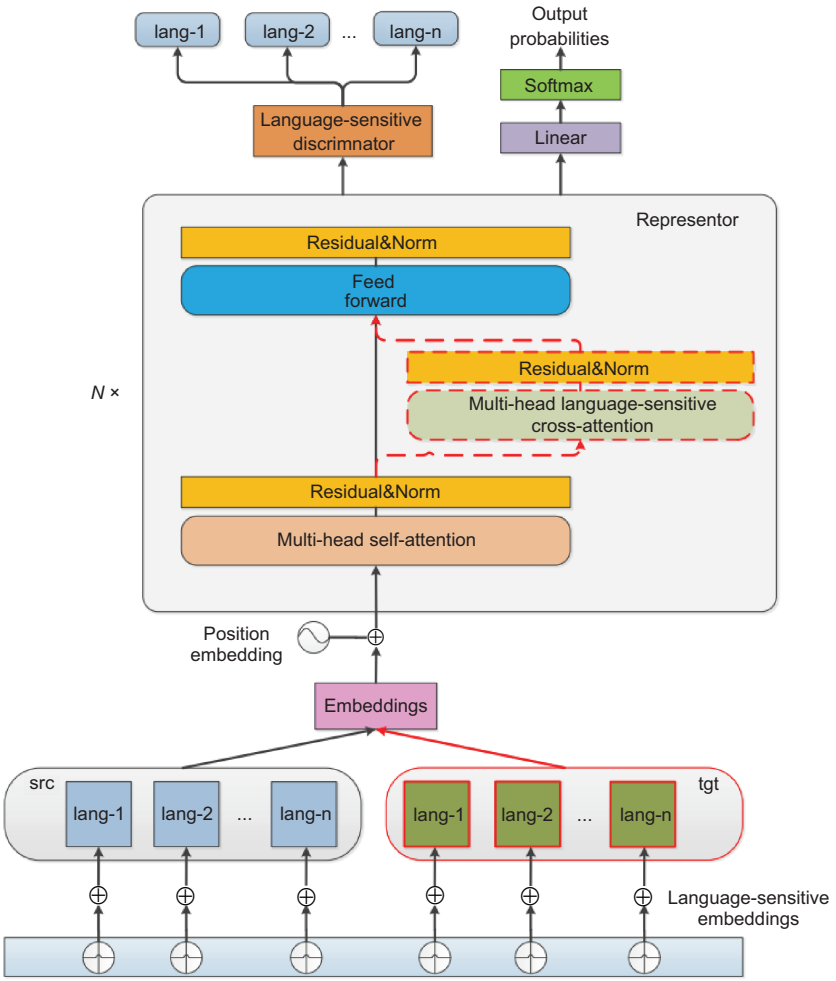

Figure 6 (Color online) Illustration of a compact and language-sensitive multilingual NMT model. The compactness is ensured by sharing parameters between encoder and decoder, denoted as representor. Language-sensitive capacity is realized by three components: language-sensitive embedding (bottom), language-sensitive cross-attention (middle) and language discriminator (top).

for this module, they employed a neural model f_{dis} on the top layer of reprensentor $\boldsymbol{h}^{\text{rep}}_{\text{top}}$, and this model outputs a language judgment score P_{lang}.

$$P_{\text{lang}} = \text{softmax}(W_{\text{dis}} \times f_{\text{dis}}(\boldsymbol{h}^{\text{rep}}_{\text{top}}) + b_{\text{dis}}). \tag{12}$$

Combining the above four ideas together, they showed through extensive experiments that the new method significantly improves multilingual NMT on one-to-many, many-to-many and zero-shot scenarios, outperforming bilingual counterparts in most cases. It indicates that low-resource language translation can greatly benefit from this kind of multilingual NMT, and so do zero-resource language translations.

6.2 Semi-supervised neural machine translation

Semi-supervised neural machine translation is a paradigm which aims at building a good NMT system with limited bilingual training data $\mathcal{D} = \{\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)}\}_{m=1}^{M}$ plus massive source monolingual data $\mathcal{D}_x = \{\boldsymbol{x}^{(l_x)}\}_{l_x=1}^{L_x}$ or target monolingual data $\mathcal{D}_y = \{\boldsymbol{y}^{(l_y)}\}_{l_y=1}^{L_y}$ or both.

Monolingual data play a very important role in SMT where the target-side monolingual corpus is leveraged to train a language model (LM) to measure the fluency of the translation candidates during decoding [4,5,79]. Using monolingual data as a language model in NMT is not trivial since it needs to modify the architecture of the NMT model. Refs. [80, 81] integrated NMT with LM by combining hidden states of both models, making the model much complicated.

As for leveraging the target-side monolingual data, back-translation (BT) proposed by ref. [82] may be one of the best solutions up to now. BT is easy and simple to use since it is model agnostic to the NMT framework [83, 84]. It only requires to train a target-to-source translation system to translate the target-side monolingual sentences back into source language. The source translation and its corresponding target sentence are paired as pseudo bitexts which combined to-

gether with original bilingual training data to train the source-to-target NMT system. It has been proven to be particularly useful in low-resource translation [85]. Ref. [84] conducted a deep analysis to understand BT and investigate various methods for synthetic source sentence generation. Ref. [86] proposed to measure the confidence level of synthetic bilingual sentences so as to filter the noise.

In order to utilize the source-side monolingual data, ref. [87] proposed two methods: forward translation and multi-task learning. Forward translation is similar to BT, and the multi-task learning method performs source-to-target translation task and source sentence reordering task by sharing the same encoder.

Many researchers resort to use both side monolingual data in NMT at the same time [88–91]. We summarize two methods in Figure 7: the auto-encoder based semi-supervised learning method [88] and the dual learning method [89]. For a source-side monolingual sentence x, ref. [88] employed source-to-target translation as encoder to generate latent variable y and leverage target-to-source translation as decoder to reconstruct the input leading to x'. They optimize the parameters by maximizing the reconstruction probability as shown in Figure 7(a). The target-side monolingual data are used in a symmetric way. Figure 7(b) shows the objective function for the dual learning method. Ref. [89] treated source-to-target translation as the primal task and target-to-source translation as the dual task. Agent A sends through the primal task a translation of the source monolingual sentence to the agent B. B is responsible to estimate the quality of the translation with a language model and the dual task. The rewards including the similarity between the input x and reconstructed one x', and two language model scores $LM(y)$, $LM(x')$, are employed to optimize the network parameters of both source-to-target and target-to-source NMT models. Similarly, the target-side monolingual data are used in a symmetric way in dual learning.

Ref. [90] introduced an iterative back-translation algorithm to exploit both source and target monolingual data with

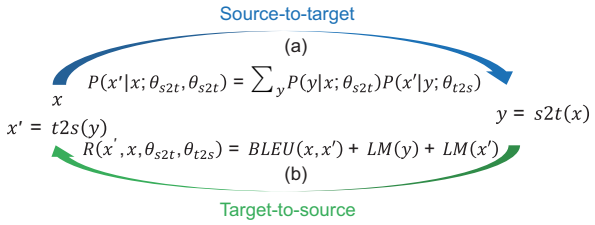

Figure 7 (Color online) Illustration of two methods exploring monolingual data. If the parameters are trained to maximize the objective function of (a), it is the auto-encoder based method. If using reward as (b) shows, it is the dual learning method. Note that this figure only demonstrates the usage of source-side monolingual data for simplicity. The use of target-side monolingual data is symmetric.

an EM optimization method. Ref. [91] proposed a mirror-generative NMT model, that explores the monolingual data by unifying the source-to-target NMT model, the target-to-source NMT model, and two language models. They showed better performance can be achieved compared to back-translation, iterative back-translation and dual learning.

6.3 Unsupervised neural machine translation

Unsupervised neural machine translation addresses a very challenging scenario in which we are required to build a NMT model using only massive source-side monolingual data $\mathcal{D}_x = \{\boldsymbol{x}^{(l_x)}\}_{l_x=1}^{L_x}$ and target-side monolingual data $\mathcal{D}_y = \{\boldsymbol{y}^{(l_y)}\}_{l_y=1}^{L_y}$.

Unsupervised machine translation can date back to the era of SMT, in which decipherment approach is employed to learn word translations from monolingual data [92–94] or bilingual phrase pairs can be extracted and their probabilities can be estimated from monolingual data [95, 96].

Since ref. [97] found that word embeddings from two languages can be mapped using some seed translation pairs, bilingual word embedding learning or bilingual lexicon induction has attracted more and more attention [98–103]. Refs. [101, 102] applied linear embedding mapping and adversarial training to learn word pair matching in the distribution level and achieve promising accuracy for similar languages.

Bilingual lexicon induction greatly motivates the study of unsupervised NMT on sentence level. And two techniques of denoising auto-encoder and back-translation make it possible for unsupervised NMT. The key idea is to find a common latent space between the two languages. Refs. [104, 105] both optimized dual tasks of source-to-target and target-to-source translation. Ref. [104] employed shared encoder to force two languages into a same semantic space, and two language-dependent decoders. In contrast, ref. [105] ensured the two languages share the same encoder and decoder, relying on an identifier to indicate specific language similar to single-model based multilingual NMT [71]. The architecture and training objective functions are illustrated in Figure 8.

The top in Figure 8 shows the use of denoising auto-encoder. The encoder encodes a noisy version of the input x into hidden representation z_{src} which is used to reconstruct the input with the decoder. The distance (auto-encoder loss $\mathcal{L}_{\text{auto}}$) between the reconstruction x' and the input x should be as small as possible. To guarantee source and target languages share the same semantic space, an adversarial loss $\mathcal{L}_{\text{adv}}$ is introduced to fool the language identifier.

The bottom in Figure 8 illustrates the use of back-translation. A target sentence y is first back-translated into x^* using an old target-to-source NMT model (the one optimized

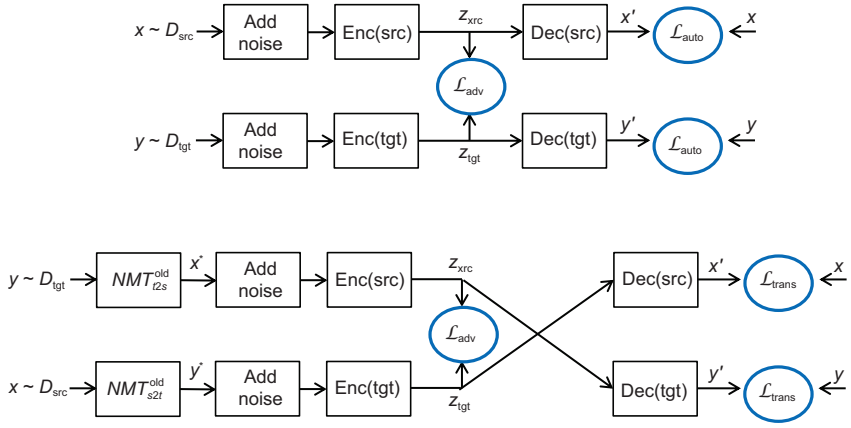

Figure 8 (Color online) Architecture of the unsupervised NMT model. The top shows denosing auto-encoder that aims at reconstructing the same language input. The bottom demonstrates back-translation which attempts to reconstruct the input in the other language using back-translation (target-to-source) and forward translation (source-to-target). The auto-encoder loss $\mathcal{L}_{\text{auto}}$, the translation loss $\mathcal{L}_{\text{trans}}$ and the language adversarial loss $\mathcal{L}_{\text{adv}}$ are used together to optimize the dual NMT models.

in previous iteration, and the initial model is the word-by-word translation model based on bilingual word induction). Then, the noisy version of the translation x^* is encoded into z_{src} which is then translated back into target sentence y'. The new NMT model (encoder and decoder) is optimized to minimize the translation loss $\mathcal{L}_{\text{trans}}$ which is the distance between the translation y' and the original target input y. Similarly, an adversarial loss is employed in the encoder module. This process iterates until convergence of the algorithm. Finally, the encoder and decoder can be applied to perform dual translation tasks.

Ref. [106] argued that sharing some layers of encoder and decoder while making others language-specific could improve the performance of unsupervised NMT. Ref. [107] further combined the NMT and SMT to improve the unsupervised translation quality. Most recently, refs. [108–110] resorted to pre-training techniques to enhance the unsupervised NMT model. For example, ref. [108] proposed a cross-lingual language model pre-training method under BERT framework [111]. Then, two pre-trained cross-lingual language models are employed as the encoder and decoder respectively to perform translation.

7 Multimodal neural machine translation

We know that humans communicate with each other in a multimodal environment in which we see, hear, smell and so on. Naturally, it is ideal to perform machine translation with the help of texts, speeches and images. Unfortunately, video corpora containing parallel texts, speech and images for machine translation are not publicly available currently. Recently, IWSLT-2020[4] organized the first evaluation on vedio translation in which annotated video data are only available for validation and test sets.

Translation for paired image-text, offline speech-to-text translation and simultaneous translation have become increasingly popular in recent years.

7.1 Image-text translation

Given an image and its text description as source language, the task of image-text translation aims at translating the description in source language into the target language, where the translation process can be supported by information from the paired image. It is a task requiring the integration of natural language processing and computer vision. WMT[5] organized the first evaluation task on image-text translation (they call it multimodal translation) in 2016 and also released the widely-used dataset Multi30K consisting of about 30K images each of which has an English description and translations in German, French and Czech[6]. Several effective models have been proposed since then. These methods mainly differ in the usage of the image information and we mainly discuss four of them in this section.

Ref. [112] proposed to encode the image into one distributed vector representation or a sequence of vector representations using convolutional neural networks. Then, they padded the vector representations together with the sentence as the final input for the NMT model which does not need to be modified for adaptation. The core idea is that they did not distinguish images from texts in the model design.

Ref. [113] presented a doubly-attentive decoder for the image-text translation task. The major difference from ref.

4) http://iwslt.org/doku.php?id=evaluation
5) https://www.statmt.org/wmt16/multimodal-task.html
6) https://github.com/multi30k/dataset

[112] is that they design textual encoder and visual encoder respectively, and employ two separate attention models to balance the contribution of text and image during prediction at each time-step.

Ref. [114] introduced a multi-task learning framework to perform image-text translation. They believe that one can imagine the image given the source language sentence. Based on this assumption, they use one encoder and two decoders in a multi-task learning framework. The encoder first encodes the source sentence into distributed semantic representations. One decoder generates the target language sentence from the source-side representations. The other decoder is required to reconstruct the given image. It is easy to see that the images are only employed in the training stage but are not required during testing. From this perspective, the multi-task learning framework can be applicable in more practical scenarios.

Ref. [115] further proposed a latent variable model for image-text translation. Different from previous methods, they designed a generative model in which a latent variable is in charge of generating the source and target language sentences, and the image as well.

Figure 9 illustrates the comparison between the doubly-attentive model and the image imagination model. Suppose the paired training data are $D = \{(\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)}, \mathrm{IM}^{(m)})\}_{m=1}^{M}$ where IM denotes image. The objective function of the doubly-attentive model can be formulated as follows:

$$\mathcal{L}(\theta) = \sum_{m=1}^{M} \log P(\boldsymbol{y}^{(m)} | \boldsymbol{x}^{(m)}, \mathrm{IM}^{(m)}; \theta). \tag{13}$$

In contrast, the image imagination model has the following objective function which includes two parts, one for text translation and the other for image imagination.

$$\mathcal{L}(\theta) = \sum_{m=1}^{M} \Big(\log P(\boldsymbol{y}^{(m)} | \boldsymbol{x}^{(m)}; \theta) + \log P(\mathrm{IM}^{(m)} | \boldsymbol{x}^{(m)}; \theta) \Big). \tag{14}$$

All the above methods are proven to significantly improve the translation quality. But it remains a natural question that when and how does the image help the text translation. Ref. [116] conducted a detailed error analysis when translating both visual and non-visual terms. They find that almost all kinds of translation errors (not only the terms having strong visual connections) have decreased after using image as the additional context.

Alternatively, ref. [117] attempted to answer when the visual information is needed in the image-text translation. They designed an input degradation method to mask crucial information in the source sentence (e.g., masking color words or entities) in order to see whether the paired image would make up the missing information during translation. They find that visual information of the image can be helpful when it is complementary rather than redundant to the text modality.

7.2 Offline speech-to-text translation

Speech-to-text translation abbreviated as speech translation (ST) is a task that automatically converts the speech in the source language (e.g., English) into the text in the target language (e.g., Chinese). Offline speech translation indicates that the complete speech (e.g., a sentence or a fragment in a time interval) is given before we begin translating. Typically, ST is accomplished with two cascaded components. Source language speech is first transcribed into the source language text using an automatic speech recognition (ASR) system. Then, the transcription is translated into target language text with a neural machine translation system. It is still the mainstream approach to ST in real applications. In this kind of paradigm, ASR and NMT are not coupled and can be optimized independently.

Nevertheless, the pipeline method has two disadvantages. On one hand, the errors propagate through the pipeline and the ASR errors are difficult to make up during translation. On the other hand, the efficiency is limited due to the two-phase process. Ref. [118] believed in early years that end-to-end speech translation is possible with the development of memory, computational capacity and representation

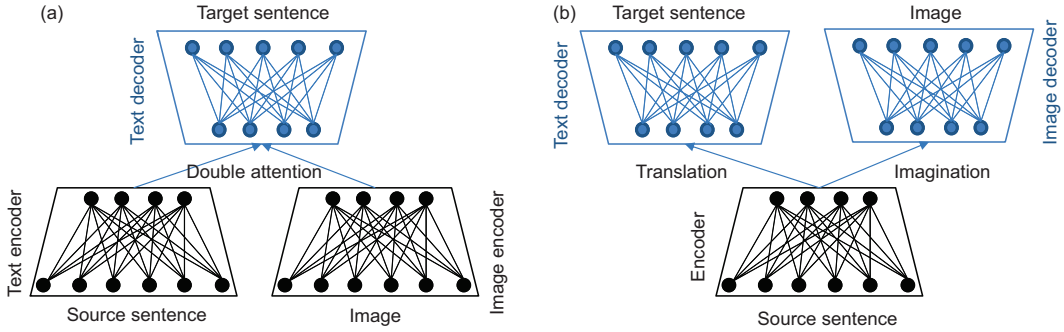

Figure 9 (Color online) Comparison between the doubly-attentive model and the image imagination model for image-text translation. In doubly attentive model, the image is encoded as an additional input feature. While in image imagination model, the image is decoded output from the source sentence.

models. Deep learning based on distributed representations facilitates the end-to-end modeling for speech translation. Ref. [119] presented an end-to-end model without using any source language transcriptions under an encoder-decoder framework. Different from the pipeline paradigm, the end-to-end model should be optimized on the training data consisting of instances (source speech, target text). We list some of the recently used datasets in Table 1 [120–123], including IWSLT[7], Augmented Librispeech[8], Fisher and Callhome[9], MuST-C[10] and TED-Corpus[11].

It is easy to find from the Table 1 that the training data for end-to-end ST are much less than that in ASR and MT. Accordingly, most of recent studies focus on fully utilizing the data or models of ASR and NMT to boost the performance of ST. Multi-task learning [124–126], knowledge distillation [127,128] and pre-training [129,130] are three main research directions.

In the multi-task learning framework, the ST model is jointly trained with the ASR and MT models. Since the ASR and MT models are optimized on massive training data, the ST model can be substantially improved through sharing encoder with the ASR model and decoder with the MT model. Ref. [124] showed that great improvements can be achieved under multi-task learning. While ref. [126] demonstrated that multi-task learning could also accelerate the convergence in addition to better translation quality.

In contrast to the multi-task learning framework, the pre-training method first pre-trains an ASR model or an MT model, then the encoder of ASR or the decoder of MT can be utilized to directly initialize the components of the ST model. Ref. [129] attempted to pre-train ST model with the ASR data to promote the acoustic model and showed that pre-training a speech encoder on one language can boost the translation quality of ST on a different source language. To further bridge the gap between pre-training and fine-tuning, ref. [130] only pre-trained the ASR encoder to maximize connectionist temporal classification (CTC) objective function [131]. Then, they share the projection matrix between the CTC classification layer for ASR and the word embeddings. The text sentence in MT is converted into the same length as the CTC output sequence of the ASR model. By doing this, the ASR encoder and the MT encoder will be consistent in length and semantic representations. Therefore, the pre-trained encoder and attention in the MT model can be used in ST in addition to the ASR encoder and the MT decoder.

Different from the multi-task learning framework and the pre-training methods which attempt to share network parameters among ASR, ST and MT, the knowledge distillation methods consider the ST model as a student and make it learn from the teacher (e.g. the MT model). Ref. [128] proposed the knowledge distillation model as shown in Figure 10. Given the training data of ST $D = \{(\boldsymbol{s}^{(m)}, \boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)})\}_{m=1}^{M}$, where $\boldsymbol{s}$ denotes the speech segment, $\boldsymbol{x}$ is the transcription in source language and $\boldsymbol{y}$ is the translation text in target language.

The objective function for ST is similar to MT and the only difference is that the input is speech segment rather than a text sentence.

$$\mathcal{L}_{ST}(\theta) = -\sum_{(\boldsymbol{s},\boldsymbol{y})\in D} \log P(\boldsymbol{y}^{(m)}|\boldsymbol{s}^{(m)};\theta), \tag{15}$$

$$\log P(\boldsymbol{y}|\boldsymbol{s},\theta) = \sum_{i=0}^{I}\sum_{v=1}^{|V|} \mathbb{I}(y_i = v)\log P(y_i|\boldsymbol{s},\boldsymbol{y}_{<i},\theta), \tag{16}$$

where $|V|$ denotes the vocabulary size of the target language and $\mathbb{I}(y_i = v)$ is the indication function which indicates whether the i-th output token y_i happens to be the ground truth.

Given the MT teacher model pre-trained on large-scale data, it can be used to force decode the pair $(\boldsymbol{x},\boldsymbol{y})$ from the triple $(\boldsymbol{s},\boldsymbol{x},\boldsymbol{y})$ and will obtain a probability distribution for each target word y_i: $Q(y_i|\boldsymbol{x},\boldsymbol{y}_{<i};\theta_{\mathrm{MT}})$. Then, the knowledge distillation loss can be written as follows:

Table 1 some datasets used in the end-to-end ST. En, De, Fr, Es, It, Nl, Pt, Ro, Ru, Zh and Ja denote English, German, French, Spanish, Italian, Dutch, Portuguese, Romanian, Russian, Chinese and Japanese, respectively

Corpus name	Source language	Target language	Hours	Sentents
IWSLT [120]	En	De	273	171,121
Augmented Librispeech	En	Fr	236	131,395
Fisher and Callhome [121]	En	Es	160	153,899
MuST-C [122]	En	De, Es, Fr, It, Nl, Pt, Ro, Ru	385-504	4.0M-5.3M
TED-Corpus [123]	En	De, Fr, Zh, Ja	520	235K-299K

7) http://i13pc106.ira.uka.de/ mmueller/iwslt-corpus.zip
8) https://persyval-platform.univ-grenoble-alpes.fr/DS91/detaildataset
9) https://github.com/joshua-decoder/fisher-callhome-corpus
10) https://mustc.fbk.eu
11) https://drive.google.com/drive/folders/1sFe6Qht4vGD49vs7_gbrNEOsLPOX9VIn?usp=sharing

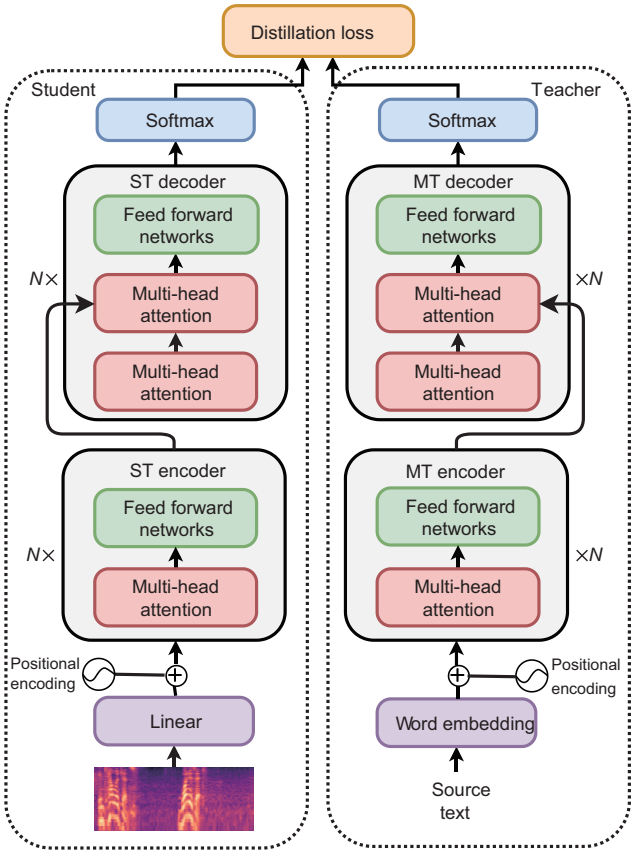

Figure 10 (Color online) The illustration of the knowledge distillation model for ST. The right part is an MT model, which is a teacher. The left part is the ST model which is the student. The input of the ST model is raw speech and the input of the MT model is the transcription of the speech. The distillation loss in the top part makes the student model learn output probability distributions from the teacher model (mimic the behavior of the teacher).

$$\mathcal{L}_{KD}(\theta) = -\sum_{(\boldsymbol{x},\boldsymbol{y})\in D}\sum_{i=0}^{I}\sum_{v=1}^{|V|} Q(y_i|\boldsymbol{x},\boldsymbol{y}_{<i};\theta_{\mathrm{MT}})\log P(y_i|\boldsymbol{x},\boldsymbol{y}_{<i},\theta). \tag{17}$$

The final ST model can be trained by optimizing both of the log-likelihood loss $\mathcal{L}_{\mathrm{ST}}(\theta)$ and the knowledge distillation loss $\mathcal{L}_{\mathrm{KD}}(\theta)$.

In order to fully explore the integration of ASR and ST, ref. [123] further proposed an interactive model in which ASR and MT perform synchronous decoding. As shown in Figure 11, the dynamic outputs of each model can be used as the context to improve the prediction of the other model. Through interaction, the quality of both models can be significantly improved while keeping the efficiency as much as possible.

12) https://autosimtrans.github.io/shared
13) http://iwslt.org/doku.php?id=simultaneous_translation

7.3 Simultaneous machine translation

Simultaneous machine translation (SimMT) aims at translating concurrently with the source-language speaker speaking. It addresses the problem where we need to incrementally produce the translation while the source-language speech is being received. This technology is very helpful for live events and real-time video-call translation. Recently, Baidu and Facebook organized the first evaluation task on SimMT in ACL-2020[12] and IWSLT-2020[13] respectively.

Obviously, the methods of offline speech translation introduced in the Sect. 7.2 cannot be applicable in these scenarios, since the latency must be intolerable if translation begins after speakers complete their utterance. Thus, balancing between latency and quality becomes the key challenge for the SimMT system. If it translates before the necessary information arrives, the quality will decrease. However, the delay will be unnecessary if it waits for too much source-language contents.

Refs. [132, 133] proposed to directly perform simultaneous speech-to-text translation, in which the model is required to generate the target-language translation from the incrementally incoming foreign speech. In contrast, more research work focus on the simultaneous text-to-text translation where they assume that the transcriptions are correct [134–143]. This article mainly introduces the latter methods. All of the methods address the same strategy (also known as policy) that when to read an input word from the source language and when to write an output word in target language, namely when to wait and when to translate.

In general, the policies can be categorized into two bins. One is fixed-latency policies [138, 139], such as wait-k policy. The other is adaptive policies [135, 136, 140–143].

The wait-k policy proposed by ref. [139] is proven simple but effective. Just as shown in the middle part of Figure 12, the wait-k policy starts to translate after reading the first k source words. Then, it alternates between generating a target-language word and reading a new source-language word, until it meets the end of the source sentence. Accordingly, the probability of a target word y_i is conditioned on the history predictions $\boldsymbol{y}_{<i}$ and the prefix of the source sentence $\boldsymbol{x}_{<i+k}$: $P(y_i|\boldsymbol{y}_{<i},\boldsymbol{x}_{<i+k};\theta)$. The probability of the whole target sentence becomes

$$P(\boldsymbol{y}|\boldsymbol{x};\theta) = \prod_{i=0}^{I} P(y_i|\boldsymbol{y}_{<i},\boldsymbol{x}_{<i+k};\theta). \tag{18}$$

In contrast to previous sequence-to-sequence NMT training paradigm, ref. [139] designed a prefix-to-prefix training

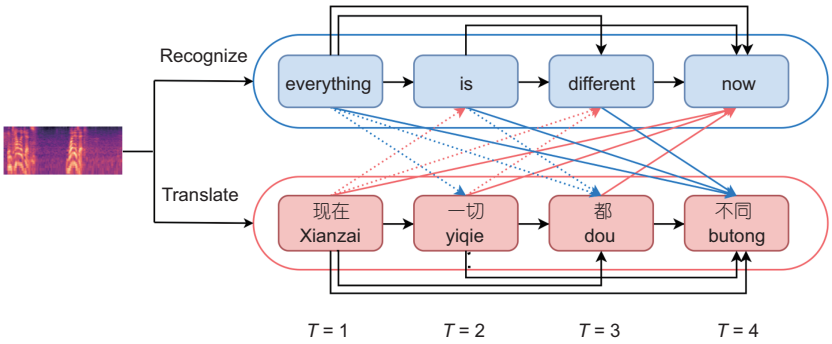

Figure 11 (Color online) The demonstration of the interactive model for both ASR and ST. Taking $T = 2$ as an example, the transcription "everything" of the ASR model can be helpful to predict the Chinese translation at $T = 2$. Likewise, the translation at time step $T = 1$ is also beneficial to generate the transcriptions of the ASR model in the future time steps.

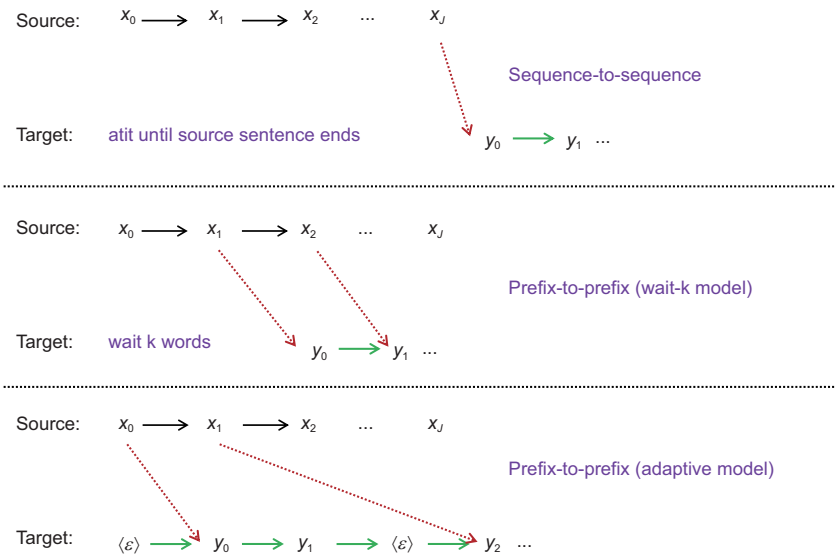

Figure 12 (Color online) The illustration of three policies for simultaneous machine translation. The top part is the conventional sequence-to-sequence MT model which begins translation after seeing the whole source sentence. The middle one demonstrates the wait-k policy which waits for k words before translation. The bottom part shows an example of the adaptive policy that predicts an output token at each time step. If the output is a special token $\langle\varepsilon\rangle$, it indicates reading one more source word.

style to best explore the wait-k policy. If Transformer is employed as the basic architecture, prefix-to-prefix training algorithm only needs a slight modification. The key difference from Transformer is that prefix-to-prefix model conditions on the first $i + k$ rather than all source words at each time-step i. It can be easily accomplished by applying the masked self-attention during encoding the source sentence. In that case, each source word is constrained to only attend to its predecessors and the hidden semantic representation of the $i + k$-th position will summarize the semantics of the prefix $\boldsymbol{x}_{<i+k}$.

However, the wait-k policy is a fixed-latency model and it is difficult to decide k for different sentences, domains and languages. Thus, adaptive policy is more appealing. Early attempts for adaptive policy are based on reinforcement learning (RL) method. For example, ref. [136] presented a two-stage model that employs the pre-trained sentence-based NMT as the base model. On top of the base model, read or translate actions determine whether to receive a new source word or output a target word. These actions are trained using the RL method by fixing the base NMT model.

Differently, ref. [141] proposed an end-to-end simMT model for adaptive policies. They first add a special delay token $\langle\varepsilon\rangle$ into the target-language vocabulary. As shown in the bottom part of Figure 12, if the model predicts $\langle\varepsilon\rangle$, it needs to receive a new source word. To train the adaptive policy model, they design dynamic action oracles with aggressive and conservative bounds as the expert policy for imitation learning. Suppose the prefix pair is $(\boldsymbol{s}, \boldsymbol{t})$. Then, the dynamic action oracle can be defined as follows:

$$\Pi^{\star}_{\boldsymbol{x},\boldsymbol{y},\alpha,\beta}(\boldsymbol{s},\boldsymbol{t}) = \begin{cases} \{\langle\varepsilon\rangle\}, & \text{if } \boldsymbol{s} \neq \boldsymbol{x} \text{ and } |\boldsymbol{s}| - |\boldsymbol{t}| \leq \alpha, \\ \{y_{|\boldsymbol{t}|+1}\}, & \text{if } \boldsymbol{t} \neq \boldsymbol{y} \text{ and } |\boldsymbol{s}| - |\boldsymbol{t}| \geq \beta, \\ \{\langle\varepsilon\rangle, y_{|\boldsymbol{t}|+1}\}, & \text{otherwise}, \end{cases}$$

where α and β are hyper-parameters, denoting aggressive and conservative bounds respectively. $|\boldsymbol{s}| - |\boldsymbol{t}|$ calculates the distance between two prefixes. That is to say if the current target prefix $\boldsymbol{t}$ is no more than α words behind the source prefix $\boldsymbol{s}$, we can read a new source word. If $\boldsymbol{t}$ is shorter than $\boldsymbol{s}$ with more than β words, we generate the next target prediction.

8 Discussion and future research tasks

8.1 NMT vs. human

We can see from Sects. 4–7 that great progresses have been achieved in neural machine translation. Naturally, we may wonder whether current strong NMT systems could perform on par with or better than human translators. Exciting news were reported in 2018 by ref. [12] that they achieved human-machine parity on Chinese-to-English news translation and they found no significant difference of human ratings between their MT outputs and professional human translations. Moreover, the best English-to-Czech system submitted to WMT 2018 by ref. [144] was also found to perform significantly better than the human-generated reference translations [145]. It is encouraging that NMT can achieve very good translations in some specific scenarios and it seems that NMT has achieved the human-level translation quality.

However, we cannot be too optimistic since the MT technology is far from satisfactory. On one hand, the comparisons were conducted only on news domain in specific language pairs where massive parallel corpora are available. In practice, NMT performs quite poorly in many domains and language pairs, especially for the low-resource scenarios such as Chinese-Hindi translation. On the other hand, the evaluation methods on the assessment of human-machine parity conducted by ref. [12] should be much improved as pointed out by ref. [146]. According to the comprehensive investigations conducted by ref. [146], human translations are much preferred over MT outputs if using better rating techniques, such as choosing professional translators as raters, evaluating documents rather than individual sentences and utilizing original source texts instead of source texts translated from target language. Current NMT systems still suffer from serious translation errors of mistranslated words or named entities, omissions and wrong word order. Obviously, there is much room for NMT to improve and we suggest some potential research directions in the next section.

8.2 Future research tasks

In this section, we discuss some potential research directions for neural machine translation.

(1) Effective document-level translation and evaluation

It is well known that document translation is important and the current research results are not so good. It remains unclear what is the best scope of the document context needed to translate a sentence. It is still a question whether it is reasonable to accomplish document translation by translating the sentences from first to last. Maybe translation based on sentence group is a worthy research topic which models many-to-many translation. In addition, document-level evaluation is as important as the document-level MT methods, and it serves as a booster of MT technology. Ref. [12] argued that MT can achieve human parity in Chinese-to-English translation on specific news tests if evaluating sentence by sentence. However, as we discussed in the previous section that [146,147] demonstrated a stronger preference for human over MT when evaluating on document-level rather than sentence-level. Therefore, how to automatically evaluate the quality of document translation besides BLEU [148] is an open question although some researchers introduce several test sets to investigate some specific discourse phenomena [149].

(2) Efficient NMT inference

People prefer the model with both high accuracy and efficiency. Despite of remarkable speedup, the quality degradation caused by non-atuoregressive NMT is intolerable in most cases. Improving the fertility model, word ordering of decoder input and dependency of the output will be worthy of a good study to make NAT close to AT in translation quality. Synchronous bidirectional decoding deserves deeper investigation due to good modeling of history and future contexts. Moreover, several researchers start to design decoding algorithm with free order [150–152] and it may be a good way to study the nature of human language generation.

(3) Making full use of multilinguality and monolingual data

Low-resource translation is always a hot research topic since most of natural languages are lack of abundant annotated bilingual data. The potential of multilingual NMT is not fully explored and some questions remain open. For example, how to deal with data unbalance problem which is very common in multilingual NMT? How to build a good incremental multilingual NMT model for incoming new languages? Semi-supervised NMT is more practical in real applications but the effective back-translation algorithm is very time consuming. It deserves to design a much efficient semi-supervised NMT model for easy deployment. Deeply integrating pre-trained method with NMT may lead to promising improvement in the semi-supervised or unsupervised learning framework and refs. [153,154] have already shown some good improvements. The achievements of unsupervised MT in similar language pairs (e.g., English-German and English-French) make us very optimistic. However, ref. [155] showed

that unsupervised MT performs poorly on distant language pairs, obtaining no more than 10 BLEU scores in most cases. Obviously, it is challenging to design better unsupervised MT models on distant language pairs.

(4) Better exploiting multimodality in NMT

In multimodal neural machine translation, it remains an open problem when and how to make full use of different modalities. The image-text translation only translates the image captions and is hard to be widely used in practice. It is a good research topic to find an appropriate scenario where images are indispensible during translation. In speech translation, despite of big improvement, the end-to-end framework currently cannot perform on par with the cascaded method in many cases, especially when the training data are limited [123]. In addition to enlarging the training data, closing the gap between different semantic spaces of ST, ASR and MT is worthy of further exploration. As for simultaneous translation, it is still on the early stage of research and many practical issues such as repetition and correction in speech are unexplored. Moreover, combining summarization and translation may be a good research direction that provides the audiences the gist of the speaker's speech with low latency.

(5) NMT with background modeling

In many cases, machine translation is not about texts, speeches and images, but is highly related to culture, environment, history and etc. Accordingly, this kind of background information should be well captured by a novel model which guides NMT to generate translations in line with the background.

(6) Incorporating prior knowledge into NMT

Note that some research topics are not mentioned in this article due to space limit. For example, it is a difficult and practical issue how to integrate prior knowledge (e.g. alignment, bilingual lexicon, phrase table and knowledge graphs) into the NMT framework. Since it is still unclear how to bridge discrete symbol based knowledge and distributed representation based NMT, it remains an interesting and important direction to explore although some progress has been achieved [156–164].

(7) Better domain adaption models

Domain adaptation has been always a hot research topic and attracts attentions from many researchers [165–172]. Different from methods used in SMT, domain adaptation in NMT is usually highly related with parameter fine-tuning. It remains a challenge how to address the problem of unknown test domain and out-of-domain term translations.

(8) Bridging the gap between training and inference

The inconsistency between training and inference (or evaluation) is a critical problem in most sequence generation tasks in addition to neural machine translation. This problem is well addressed in the community of machine translation [173, 174] but it is still worthy of exploring especially on the efficiency of the methods.

(9) Designing explainable and robust NMT

So far, the NMT model is still a black box and it is very risky to use it in many scenarios in which we have to know how and why the translation result is obtained. Ref. [175] attempted to visualize the contribution of each input to the output translation. Nevertheless, it will be great to deeply investigate the explanation of the NMT models or design an explainable MT architecture. Furthermore, current NMT systems are easy to attack through perturbing the input. Ref. [176, 177] presented novel robust NMT models to handle noisy inputs. However, the real input noise is too difficult to anticipate and it still remains a big challenge to design robust NMT models which are immune to real noise.

(10) New NMT architectures

Finally, designing better NMT architectures beyond Transformer must be very exciting to explore despite of the difficulty.

This work was supported by the National Natural Science Foundation of China (Grant Nos. U1836221 and 61673380), and the Beijing Municipal Science and Technology Project (Grant No. Z181100008918017).

1 Weaver W. Translation. Machine Trans Languages, 1955, 14: 15–23
2 Nagao M. A framework of a mechanical translation between japanese and english by analogy principle. In: Proceedings of the International NATO Symposium on Artificial and Human Intelligence. Lyon, 1984. 173–180
3 Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: Parameter estimation. Comput Linguist, 1993. 19: 263–311
4 Koehn P, Och F J, Marcu D. Statistical phrase-based translation. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Edmonton, 2003. 48–54
5 Chiang D. A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Ann Arbor, 2005. 263–270
6 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of the Conference on Neural Information Processing Systems. Montreal, 2014. 3104–3112
7 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations. San Diego, 2015
8 Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning. In: Proceedings of the International Conference on Machine Learning. Sydney, 2017. 1243–1252
9 Vawani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the Conference on Neural Information Processing Systems. Long Beach, 2017. 5998–6008
10 Junczys-Dowmunt M, Dwojak T, Hoang H. Is neural machine translation ready for deployment? A case study on 30 translation directions. ArXiv: 1610.01108
11 Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv: 1609.08144

12 Hassan H, Aue A, Chen C, et al. Achieving human parity on automatic chinese to english news translation. ArXiv: 1803.05567
13 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, 2016. 1715–1725
14 Chen M X, Firat O, Bapna A, et al. The best of both worlds: Combining recent advances in neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018. 76–86
15 Wang Q, Li B, Xiao T, et al. Learning deep transformer models for machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 1810–1822
16 Zhang B, Titov I, Sennrich R. Improving deep transformer with depth-scaled initialization and merged attention. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 898–909
17 Li Y, Wang Q, Xiao T, et al. Neural machine translation with joint representation. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, 2020. 8285–8292
18 Wu F, Fan A, Baevski A, et al. Pay less attention with lightweight and dynamic convolutions. In: Proceedings of the International Conference on Learning Representations. New Orleans, 2019
19 So D R, Liang C, Le Q V. The evolved transformer. In: Proceedings of the International Conference on Machine Learning. Long Beach, 2019. 5877–5886
20 Lu Y, Li Z, He D, et al. Understanding and improving transformer from a multi-particle dynamic system point of view. ArXiv: 1906.02762
21 Zhang J, Zong C. Deep neural networks in machine translation: An overview. IEEE Intell Syst, 2015, 30: 16
22 Liu Y, Zhang J. Deep learning in machine translation. In: Deep Learning in Natural Language Processing. Singapore: Springer, 2018. 147–183
23 Koehn P, Knowles R. Six challenges for neural machine translation. ArXiv: 1706.03872
24 Zhang J, Luan H, Sun M, et al. Improving the transformer translation model with document-level context. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 533–542
25 Xiong H, He Z, Wu H, et al. Modeling coherence for discourse neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, 2019. 7338–7345
26 Gong Z, Zhang M, Zhou G. Cache-based document-level statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, 2011. 909–919
27 Xiao T, Zhu J, Yao S, et al. Document-level consistency verification in machine translation. In: Proceedings of the 13th Machine Translation Summit. Xiamen, 2011. 131–138
28 Xiong D, Ding Y, Zhang M, et al. Lexical chain based cohesion models for document-level statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle, 2013. 1563–1573
29 Born L, Mesgar M, Strube M. Using a graph-based coherence model in document-level machine translation. In: Proceedings of the Third Workshop on Discourse in Machine Translation. Copenhagen, 2017. 26–35
30 Rios A, Tuggener D. Co-reference resolution of elided subjects and possessive pronouns in spanish-english statistical machine translation. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Valencia, 2017. 657–662
31 Kuang S, Xiong D, Luo W, et al. Modeling coherence for neural machine translation with dynamic and topic caches. In: Proceedings of the International Conference on Computational Linguistics. Santa Fe, New Mexico, 2018. 596–606
32 Tu Z, Liu Y, Shi S, et al. Learning to remember translation history with a continuous cache. Trans Assoc Comput Linguist, 2018, 6: 407–420
33 Jean S, Lauly S, Firat O, et al. Does neural machine translation benefit from larger context? ArXiv: 1704.05135
34 Wang L, Tu Z, Way A, et al. Exploiting cross-sentence context for neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017. 2826–2831
35 Voita E, Serdyukov P, Sennrich R, et al. Context-aware neural machine translation learns anaphora resolution. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018. 1264–1274
36 Miculicich L, Ram D, Pappas N, et al. Document-level neural machine translation with hierarchical attention networks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 2947–2954
37 Yang Z, Zhang J, Meng F, et al. Enhancing context modeling with a query-guided capsule network for document-level translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 1527–1537
38 Maruf S, Haffari G. Document context neural machine translation with memory networks. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018. 1275–1284
39 Maruf S, Martins A F, Haffari G. Selective attention for context-aware neural machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. 3092–3102
40 Tan X, Zhang L, Xiong D, et al. Hierarchical modeling of global context for document-level neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 1576–1585
41 Tiedemann J, Scherrer Y. Neural machine translation with extended context. In: Proceedings of the Third Workshop on Discourse in Machine Translation. Copenhagen, 2017. 82–92
42 Bawden R, Sennrich R, Birch A, et al. Evaluating discourse phenomena in neural machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, 2018. 1304–1313
43 Voita E, Sennrich R, Titov I. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 1198–1212
44 Voita E, Sennrich R, Titov I. Context-aware monolingual repair for neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 877–886
45 Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation. In: Proceedings of the International Conference on Learning Representations. Vancouver, 2018
46 Zhang X, Su J, Qin Y, et al. Asynchronous bidirectional decoding for neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, 2018. 5698–5705
47 Zhou L, Zhang J, Zong C. Synchronous bidirectional neural machine translation. Trans. Association Comput Linguist, 2019, 7: 91–105

48 Lee J, Mansimov E, Cho K. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 1173–1182
49 Wang C, Zhang J, Chen H. Semi-autoregressive neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 479–488
50 Guo J, Tan X, He D, et al. Non-autoregressive neural machine translation with enhanced decoder input. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, 2019. 3723–3730
51 Shao C, Feng Y, Zhang J, et al. Retrieving sequential information for non-autoregressive neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 3013–3024
52 Wang Y, Tian F, He D, et al. Non-autoregressive machine translation with auxiliary regularization. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, 2019. 5377–5384
53 Wei B, Wang M, Zhou H, et al. Imitation learning for non-autoregressive neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 1304–1312
54 Zheng Z, Zhou H, Huang S, et al. Modeling past and future for neural machine translation. Trans Assoc Comput Linguist, 2018, 6: 145–157
55 Zheng Z, Huang S, Tu Z, et al. Dynamic past and future for neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 931–941
56 Zhang B, Xiong D, Su J, et al. Future-aware knowledge distillation for neural machine translation. IEEE/ACM Trans Audio Speech Lang Process, 2019, 27: 2278–2287
57 Liu L, Utiyama M, Finch A, et al. Agreement on target-bidirectional neural machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, 2016. 411–416
58 Hoang C D V, Haffari G, Cohn T. Decoding as continuous optimization in neural machine translation. ArXiv: 1701.02854
59 Zhang Z, Wu S, Liu S, et al. Regularizing neural machine translation by target-bidirectional agreement. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, 2019. 443–450
60 Sennrich R, Haddow B, Birch A. Edinburgh neural machine translation systems for WMT 16. In: Proceedings of the First Conference on Machine Translation. Berlin, 2016. 371–376
61 Sennrich R, Birch A, Currey A, et al. The university of edinburgh's neural mt systems for WMT 17. In: Proceedings of the First Conference on Machine Translation. Copenhagen, 2017. 389–399
62 Su J, Zhang X, Lin Q, et al. Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding. Artificial Intelligence, 2019, 277: 103168
63 Zhou L, Zhang J, Zong C, et al. Sequence generation: From both sides to the middle. In: Proceedings of the International Joint Conference on Artificial Intelligence. Macau, 2019. 5471–5477
64 Zhang J, Zhou L, Zhao Y,et al. Synchronous bidirectional inference for neural sequence generation. Artificial Intelligence, 2020, 281: 103234
65 Wu H, Wang H. Pivot language approach for phrase-based statistical machine translation. Machine Trans, 2007, 21: 165–181
66 Cheng Y, Yang Q, Liu Y, et al. Joint training for pivot-based neural machine translation. In: Proceedings of the International Joint Conference on Artificial Intelligence. Melbourne, 2017. 3974–3980
67 Dong D, Wu H, He W, et al. Multi-task learning for multiple language translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing. Beijing, 2015. 1723–1732
68 Zoph B, Knight K. Multi-source neural translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, 2016. 30–34
69 Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, 2016. 866–875
70 Firat O, Cho K, Sankaran B, et al. Multi-way, multilingual neural machine translation. Comput. Speech Language, 2017, 45: 236–252
71 Johnson M, Schuster M, Le Q V, et al. Googles multilingual neural machine translation system: Enabling zero-shot translation. TACL, 2017, 5: 339–351
72 Blackwood G, Ballesteros M, Ward T. Multilingual neural machine translation with task-specific attention. In: Proceedings of the International Conference on Computational Linguistics. Santa Fe, 2018. 3112–3122
73 Wang Y, Zhang J, Zhai F, et al. Three strategies to improve one-to-many multilingual translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 2955–2960
74 Platanios E A, Sachan M, Neubig G, et al. Contextual parameter generation for universal neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 425–435
75 Tan X, Chen J, He D, et al. Multilingual neural machine translation with language clustering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 963–973
76 Wang Y, Zhang J, Zhou L, et al. Synchronously generating two languages with interactive decoding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 3341–3346
77 Wang Y, Zhou L, Zhang J, et al. A compact and language-sensitive multilingual translation method. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 1213–1223
78 Press O, Wolf L. Using the output embedding to improve language models. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Valencia, 2017. 157–163
79 Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Prague, 2007. 177–180
80 Gulcehre C, Firat O, Xu K, et al. On using monolingual corpora in neural machine translation. ArXiv: 1503.03535
81 Gulcehre C, Firat O, Xu K,et al. On integrating a language model into neural machine translation. Comput Speech Language, 2017, 45: 137–148
82 Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, 2016. 86–96
83 Hoang V C D, Koehn P, Haffari G, et al. Iterative back-translation for neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. Melbourne, 2018. 18–24
84 Edunov S, Ott M, Auli M, et al. Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 489–500

85 Karakanta A, Dehdari J, van Genabith J. Neural machine translation for low-resource languages without parallel corpora. Machine Translation, 2018, 32: 167–189
86 Wang S, Liu Y, Wang C, et al. Improving back-translation with uncertainty-based confidence estimation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 791–802
87 Zhang J, Zong C. Exploiting source-side monolingual data in neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, 2016. 1535–1545
88 Cheng Y, Xu W, He Z, et al. Semi-supervised learning for neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, 2016. 1965–1974
89 He D, Xia Y, Qin T, et al. Dual learning for machine translation. In: Proceedings of the Conference on Neural Information Processing Systems, Barcelona, 2016. 820–828
90 Zhang Z, Liu S, Li M, et al. Joint training for neural machine translation models with monolingual data. In: Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, 2018. 555-562
91 Zheng Z, Zhou H, Huang S, et al. Mirror-generative neural machine translation. In: Proceedings of the International Conference on Learning Representations. Addis Ababa, 2020
92 Ravi S, Knight K. Deciphering foreign language. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, 2011. 12–21
93 Dou Q, Knight K. Large scale decipherment for out-of-domain machine translation. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, 2012. 266–275
94 Nuhn M, Mauser A, Ney H. Deciphering foreign language by combining language models and context vectors. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Jeju Island, 2012. 156–164
95 Klementiev A, Irvine A, Callison-Burch C, et al. Toward statistical machine translation without parallel corpora. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Avignon, 2012. 130–140
96 Zhang J, Zong C. Learning a phrase-based translation model from monolingual data with application to domain adaptation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Sofia, 2013. 1425–1434
97 Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation. ArXiv: 1309.4168
98 Faruqui M, Dyer C. Improving vector space word representations using multilingual correlation. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, 2014. 462–471
99 Zhang M, Liu Y, Luan H, et al. Adversarial training for unsupervised bilingual lexicon induction. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017. 1959–1970
100 Zhang M, Liu Y, Luan H, et al. Earth movers distance minimization for unsupervised bilingual lexicon induction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017. 1934–1945
101 Artetxe M, Labaka G, Agirre E. Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017. 451–462
102 Conneau A, Lample G, Ranzato M, et al. Word translation without parallel data. In: Proceedings of the International Conference on Learning Representations. Vancouver, 2018
103 Cao H, Zhao T. Point set registration for unsupervised bilingual lexicon induction. In: Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, 2018. 3991–3997
104 Artetxe M, Labaka G, Agirre E. Unsupervised statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 3632–3642
105 Lample G, Conneau A, Denoyer L, et al. Unsupervised machine translation using monolingual corpora only. In: Proceedings of the International Conference on Learning Representations. Vancouver, 2018
106 Yang Z, Chen W, Wang F, et al. Unsupervised neural machine translation with weight sharing. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018. 46–55
107 Artetxe M, Labaka G, Agirre E. An effective approach to unsupervised machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 194–203
108 Conneau A, Lample G. Cross-lingual language model pretraining. In: Proceedings of the Conference on Neural Information Processing Systems. Vancouver, 2019. 7057–7067
109 Song K, Tan X, Qin T, et al. Mass: Masked sequence to sequence pre-training for language generation. In: Proceedings of the International Conference on Machine Learning. Long Beach, 2019. 5926–5936
110 Ren S, Wu Y, Liu S, et al. Explicit cross-lingual pre-training for unsupervised machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 770–779
111 Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. 4171–4186
112 Huang P Y, Liu F, Shiang S R, et al. Attention-based multimodal neural machine translation. In: Proceedings of the First Conference on Machine Translation: Shared Task Papers. Berlin, 2016. 639–645
113 Calixto I, Liu Q, Campbell N. Doubly-attentive decoder for multimodal neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017. 1913–1924
114 Elliott D, Kádár A. Imagination improves multimodal translation. In: Proceedings of the International Joint Conference on Natural Language Processing. Taipei, 2017. 130–141
115 Calixto I, Rios M, Aziz W. Latent variable model for multi-modal translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 6392–6405
116 Calixto I, Liu Q. An error analysis for image-based multi-modal neural machine translation. Machine Translation, 2019, 33: 155–177
117 Caglayan O, Madhyastha P, Specia L, et al. Probing the need for visual context in multimodal machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. 4159–4170
118 Zong C, Wu H, Huang T, et al. Analysis on characteristics of chinese spoken language. In: Proceedings of the Natural Language Processing Pacific Rim Symposium. Beijing, 1999. 358–362
119 Bérard A, Pietquin O, Servan C, et al. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In: Proceedings of the Conference on Neural Information Processing Systems Workshop on End-to-end Learning for Speech and Audio Processing. Barcelona, 2016. 1–5
120 Jan N, Cattoni R, Sebastian S, et al. The iwslt 2018 evaluation campaign. In: Proceedings of the International Workshop on Spoken Language Translation. Bruges, 2018. 2–6
121 Post M, Kumar G, Lopez A, et al. Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation

corpus. In: Proceedings of the International Workshop on Spoken Language Translation. Heidelberg, 2013. 295–301
122 Di Gangi M A, Cattoni R, Bentivogli L, et al. Must-c: A multilingual speech translation corpus. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. 2012–2017
123 Liu Y, Zhang J, Xiong H, et al. Synchronous speech recognition and speech-to-text translation with interactive decoding. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, 2020. 8417–8424
124 Weiss R J, Chorowski J, Jaitly N, et al. Sequence-to-sequence models can directly translate foreign speech. In: Proceedings of INTERSPEECH. Stockholm, 2017. 2625–2629
125 Anastasopoulos A, Chiang D. Tied multitask learning for neural speech translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, 2018. 82–91
126 Bérard A, Besacier L, Kocabiyikoglu A C, et al. End-to-end automatic speech translation of audiobooks. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Alberta, 2018. 6224–6228
127 Jia Y, Johnson M, Macherey W, et al. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Brighton, 2019. 7180–7184
128 Liu Y, Xiong H, He Z, et al. End-to-end speech translation with knowledge distillation. In: Proceedings of INTERSPEECH. Graz, 2019. 1128–1132
129 Bansal S, Kamper H, Livescu K, et al. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. 58–68
130 Wang C, Wu Y, Liu S, et al. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, 2020. 9161–9168
131 Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning. Pittsburgh, 2006. 369–376
132 Niehues J, Nguyen T S, Cho E, et al. Dynamic transcription for low-latency speech translation. In: Proceedings of INTERSPEECH. 2016. 2513–2517
133 Niehues J, Pham N Q, Ha T L, et al. Low-latency neural speech translation. In: Proceedings of INTERSPEECH. Hyderabad, 2018. 1293–1297
134 Grissom II A, He H, Boyd-Graber J, et al. Dont until the final verb wait: Reinforcement learning for simultaneous machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, 2014. 1342–1352
135 Satija H, Pineau J. Simultaneous machine translation using deep reinforcement learning. In: Proceedings of the International Conference on Machine Learning Workshop on Abstraction in Reinforcement Learning. New York, 2016
136 Gu J, Neubig G, Cho K, et al. Learning to translate in real-time with neural machine translation. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Valencia, 2017. 1053–1062
137 Alinejad A, Siahbani M, Sarkar A. Prediction improves simultaneous neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 3022–3027
138 Dalvi F, Durrani N, Sajjad H, et al. Incremental decoding and training methods for simultaneous translation in neural machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, 2018. 493–499
139 Ma M, Huang L, Xiong H, et al. Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 3025–3036
140 Arivazhagan N, Cherry C, Macherey W, et al. Monotonic infinite lookback attention for simultaneous machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 1313–1323
141 Zheng B, Zheng R, Ma M, et al. Simpler and faster learning of adaptive policies for simultaneous translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Hong Kong, 2019. 1349–1354
142 Zheng B, Zheng R, Ma M, et al. Simultaneous translation with flexible policy via restricted imitation learning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 5816–5822
143 Arthur P, Cohn T, Haffari G. Learning coupled policies for simultaneous machine translation. ArXiv: 2002.04306
144 Popel M. Cuni transformer neural mt system for WMT 18. In: Proceedings of the Third Conference on Machine Translation. Belgium, 2018. 482–487
145 Bojar O, Federmann C, Fishel M, et al. Findings of the 2018 conference on machine translation (WMT18). In: Proceedings of the Third Conference on Machine Translation. Belgium, 2018. 272–303
146 Läubli S, Castilho S, Neubig G, et al. A set of recommendations for assessing human-machine parity in language translation. J Artif Intell Res, 2020, 67: 653–672
147 Läubli S, Sennrich R, Volk M. Has machine translation achieved human parity? A case for document-level evaluation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 4791–4796
148 Papineni K, Roukos S, Ward T, et al. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Philadelphia, 2002. 311–318
149 Müller M, Rios A, Voita E, et al. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In: Proceedings of the Third Conference on Machine Translation. Brussels, 2018. 61–72
150 Gu J, Liu Q, Cho K. Insertion-based decoding with automatically inferred generation order. Trans Association Comput Linguistics, 2019, 7: 661–676
151 Stern M, Chan W, Kiros J, et al. Insertion transformer: Flexible sequence generation via insertion operations. ArXiv: 1902.03249
152 Emelianenko D, Voita E, Serdyukov P. Sequence modeling with unconstrained generation order. In: Proceedings of the Conference on Neural Information Processing Systems. Vancouver, 2019. 7698–7709
153 Ji B, Zhang Z, Duan X, et al. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, 2020. 115–122
154 Zhu J, Xia Y, Wu L, et al. Incorporating bert into neural machine translation. In: Proceedings of the International Conference on Learning Representations. Addis Ababa, 2020
155 Leng Y, Tan X, Qin T, et al. Unsupervised pivot translation for distant languages. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 175–183
156 Zhang J, Liu Y, Luan H, et al. Prior knowledge integration for neural machine translation using posterior regularization. In: Proceedings of the Annual Meeting of the Association for Computational Linguis-

tics. Vancouver, 2017. 1514–1523
157 Tu Z, Lu Z, Liu Y, et al. Modeling coverage for neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, 2016. 76–85
158 Mi H, Sankaran B, Wang Z, et al. Coverage embedding models for neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, 2016. 955–960
159 Feng Y, Zhang S, Zhang A, et al. Memory-augmented neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Copenhagen, 2019. 1390–1399
160 Zhao Y, Wang Y, Zhang J, et al. Phrase table as recommendation memory for neural machine translation. In: Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, 2018. 4609–4615
161 Zhao Y, Zhang J, He Z, et al. Addressing troublesome words in neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018. 391–400
162 Wang X, Tu Z, Xiong D, et al. Translating phrases in neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017. 1421–1431
163 Wang X, Tu Z, Zhang M. Incorporating statistical machine translation word knowledge into neural machine translation. IEEE/ACM Trans Audio Speech Language Process, 2018, 26: 2255–2266
164 Lu Y, Zhang J, Zong C. Exploiting knowledge graph in neural machine translation. In: Proceedings of the China Workshop on Machine Translation. Fujian, 2018. 27–38
165 Luong M T, Manning C D. Stanford neural machine translation systems for spoken language domain. In: Proceedings of the International Workshop on Spoken Language Translation. Da Nang, 2015. 94–97
166 Zoph B, Yuret D, May J, et al. Transfer learning for low-resource neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, 2016. 1568–1575
167 Wang R, Utiyama M, Liu L, et al. Instance weighting for neural machine translation domain adaptation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017. 1482–1488
168 Chu C, Dabre R, Kurohashi S. An empirical comparison of simple domain adaptation methods for neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017. 385–391
169 Chu C, Wang R. A survey of domain adaptation for neural machine translation. In: Proceedings of the International Conference on Computational Linguistics. Santa Fe, 2018. 1304–1319
170 Li X, Zhang J, Zong C. One sentence one model for neural machine translation. In: Proceedings of the International Conference on Language Resources and Evaluation. Miyazaki, 2018. 910–917
171 Zhang X, Shapiro P, Kumar G, et al. Curriculum learning for domain adaptation in neural machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. 1903–1915
172 Zeng J, Liu Y, Su J, et al. Iterative dual domain adaptation for neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, 2019. 845–855
173 Shen S, Cheng Y, He Z, et al. Minimum risk training for neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, 2016. 1683–1692
174 Zhang W, Feng Y, Meng F, et al. Bridging the gap between training and inference for neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 4334–4343
175 Ding Y, Liu Y, Luan H, et al. Visualizing and understanding neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017. 1150–1159
176 Cheng Y, Tu Z, Meng F, et al. Towards robust neural machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018. 1756–1766
177 Cheng Y, Jiang L, Macherey W. Robust neural machine translation with doubly adversarial inputs. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Florence, 2019. 4324–4333