Machine Intelligence Research
www.mi-research.net
DOI: 10.1007/s11633-022-1393-5

Transformer: A General Framework from Machine Translation to Others

Yang Zhao[1,2] Jiajun Zhang[1,2] Chengqing Zong[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract: Machine translation is an important and challenging task that aims at automatically translating natural language sentences from one language into another. Recently, Transformer-based neural machine translation (NMT) has achieved great breakthroughs and has become a new mainstream method in both methodology and applications. In this article, we conduct an overview of Transformer-based NMT and its extension to other tasks. Specifically, we first introduce the framework of Transformer, discuss the main challenges in NMT and list the representative methods for each challenge. Then, the public resources and toolkits in NMT are listed. Meanwhile, the extensions of Transformer in other tasks, including the other natural language processing tasks, computer vision tasks, audio tasks and multi-modal tasks, are briefly presented. Finally, possible future research directions are suggested.

Keywords: Neural machine translation, Transformer, document neural machine translation (NMT), multimodal NMT, low-resource NMT.

Citation: Y. Zhao, J. Zhang, C. Zong. Transformer: A general framework from machine translation to others. *Machine Intelligence Research*. http://doi.org/10.1007/s11633-022-1393-5

1 Introduction

Machine translation (MT) aims at automatically translating natural language sentences using computers from one language into another. Since the first MT system was proposed, it has become one of the most important and challenging tasks in natural language processing (NLP) or even in the artificial intelligence community. With the effort of many researchers, MT has achieved remarkable progress in both methodology and applications.

With the rapid development of machine learning and the availability of large-scale parallel corpora, statistical machine translation (SMT) approaches[1, 2] appeared in the 1990s and have drawn much attention. Instead of designing the translation rules manually, SMT learns the language model and word or phrase mappings automatically from the parallel corpora. However, SMT represents the source and target sentences as symbolic and discrete tokens. Thus, the performance of SMT is far from satisfactory.

With the breakthrough of deep learning, many studies have incorporated deep neural networks into MT. Early studies are still based on the SMT framework, where deep neural networks are utilized to design new features or extract more accurate semantic representations[3]. In 2013 and 2014, end-to-end neural machine translation (NMT)[4–6] has emerged as a new paradigm and quickly replaced SMT as the mainstream approach. NMT adopts the distributed representation of sentences and utilizes a whole neural network to learn the mappings from source sentences to target sentences. In only a few years of development, the translation quality of NMT has significantly improved and exceeded that of SMT. In practice, many companies (such as Google, Microsoft and Baidu) have deployed their own online translation systems and provide users with increasingly high-quality translation services[7, 8].

From the perspective of NMT architectures, the early architectures are recurrent neural network (RNN) based NMT[4–6] and convolutional neural network (CNN)[9] based NMT models, which utilize the RNN and CNN to calculate the representation of source sentences and predict the target sentence. In 2017, a new framework, self-attention based NMT (Transformer), was proposed and sharply advanced the field of NMT[10]. At present, Transformer has become the dominant architecture for machine translation, surpassing convolutional and recurrent neural network based NMT in terms of both translation quality and training speed. Meanwhile, Transformer goes far beyond NMT and extends to other tasks, such as other natural language processing tasks, computer vision tasks, audio tasks and multimodal tasks.

In this article, we attempt to give a survey of Trans-

Review
Manuscript received on September 14, 2022; accepted on November 7, 2022
Recommended by Associate Editor Ji-Rong Wen
Colored figures are available in the online version at https://link.springer.com/journal/11633

former-based NMT, including the frameworks, the main challenges, the representative methods for each challenge and the available data and toolkits in NMT. We also briefly present the extensions of Transformer in other NLP tasks, including pre-training language models, text summarization, dialogue and knowledge graphs. Finally, possible future research directions are discussed.

The remainder of this survey is organized as follows. Section 2 introduces the encoder-decoder framework and Transformer. Section 3 lists the main challenges of NMT. Section 4 represents the representative approaches for each challenge. Section 5 shows the resources and toolkits in NMT. Section 6 briefly presents the applications of Transformer in other tasks. Section 7 introduces the current status of NMT. Section 8 suggests some potential research directions.

2 Neural machine translation

Since 2013, there have been various model architectures for NMT, such as recurrent neural network-based NMT (RNMT)[5, 6, 11], convolutional neural network-based model (ConS2S)[9, 12], and a self-attention neural network-based model (Transformer)[10]. At present, self-attention neural networks are the state-of-the-art and have been widely used. In this section, we mainly introduce the encoder-decoder framework and Transformer.

2.1 Encoder-decoder framework

Sequence-to-sequence learning with an encoder-decoder framework was first proposed by Sutskever et al.[5] and Bahdanau et al.[6]. The current NMT models still follow this encoder-decoder framework[10]. Fig. 1 shows the encoder-decoder framework. As the name suggests, the encoder transforms the source sentence $X = \{x_1, x_2, \cdots, x_m\}$ into hidden states $\boldsymbol{H} = (\boldsymbol{h}_1, \cdots, \boldsymbol{h}_i, \cdots, \boldsymbol{h}_m)$. The decoder generates the target translation $Y = \{y_1, y_2, \cdots, y_n\}$ from the hidden states $\boldsymbol{H}$. Generally, the current encoder-decoder framework consists of four basic components: the embedding layer, the encoder network, the decoder network, and the softmax layer.

Embedding layer. The embedding layer maps a discrete source sentence $X = (x_1, \cdots, x_i, \cdots, x_m)$ into continuous embeddings $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_m)$, where $\boldsymbol{x}_i$ is embedding for the i-th token. Then, embeddings are fed into the encoder network.

Encoder network. It maps the source embeddings $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_m)$ into source hidden states $\boldsymbol{H} = (\boldsymbol{h}_1, \cdots, \boldsymbol{h}_i, \cdots, \boldsymbol{h}_m)$. The encoder network can be implemented by a recurrent neural network[5, 6], convolutional neural network[9, 12] or self-attention-based neural network[10]. The encoder procedure can be represented by

$$\boldsymbol{H} = \text{Encoder}(\boldsymbol{X}). \tag{1}$$

Decoder network. The decoder network generates the target sentence $Y = \{y_1, y_2, \cdots, y_n\}$ word by word. Given already produced tokens $y_{<j} = \{y_1, y_2, \cdots, y_{j-1}\}$ at the j-th time step and the hidden states $\boldsymbol{H}$, the decoder obtains the target hidden states $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{j-1})$ by

$$\boldsymbol{Z} = \text{Decoder}(y_{<j}, \boldsymbol{H}). \tag{2}$$

Similar to the encoder, the decoder network can also be implemented by a recurrent neural network, convolutional neural network or self-attention-based neural network.

Softmax layer. Finally the hidden states $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{j-1})$ of the decoder network are fed into a softmax layer to predict each token at the jth time step. More specifically, a liner layer is first utilized to transform the hidden states $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{j-1})$ into the score for each target token. Then, a softmax layer is utilized to obtain the predicted probability of j-th token $p(y_j|y_{<j}, X)$.

Given the parallel training dataset $D = \{(X, Y)\}$, where X denotes the source sentence and Y denotes the target sentence, the network parameters of the NMT θ can be optimized by maximizing the following log-likeli-

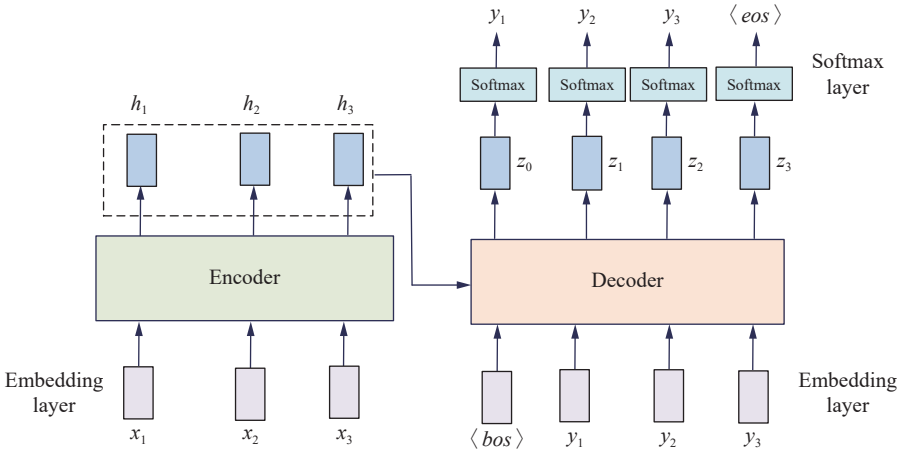


Fig. 1 Encoder-decoder framework, where encoder transforms the source sentence into hidden states and decoder generates target translation from the hidden states

hood objective function:

$$J(D,\theta) = \sum_{(X,Y)\in D} \log\ P(Y \mid X;\theta) = \sum_{(X,Y)\in D} \sum_{j=1}^{n} \log\ P(y_j \mid y_{<j}, X;\theta). \quad (3)$$

2.2 Transformer

Self-attention. Self-attention is the core component of Transformer, which can be seen as a mapping from queries Q, keys K and values V to an output. The output could attend to the information of different tokens and is computed as a weighted sum of the values V, in which the weight is determined by the queries Q and the keys K. More specifically, there are two important attention mechanisms in Transformer, i.e., scaled dot-product attention and multi-head attention. Fig. 2 shows the framework of these two attention mechanisms.

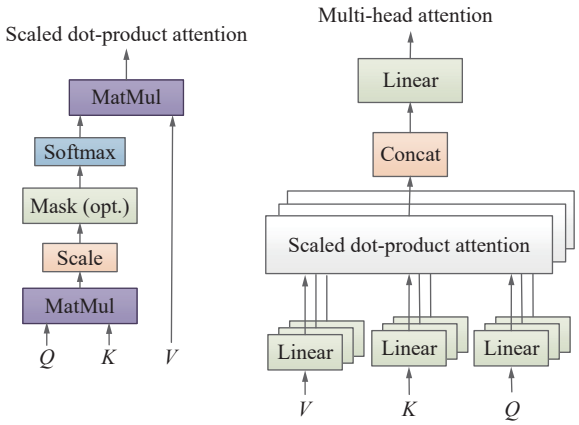


Fig. 2 The attention mechanisms in Transformer. Scaled dot-product attention (left) and multi-head attention (right) in Transformer[10].

1) Scaled dot-product attention: Given the queries Q, keys K (whose dimension is denoted by d_k) and values V, it first computes the dot products of the queries Q with all keys K, and divides the results by $\sqrt{d_k}$. Then a softmax function is utilized to obtain the weights. Finally, the attention output is calculated by multiplying the weights and the values. Formally, the procedure can be depicted as follows:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V. \quad (4)$$

2) Multi-head attention: Multi-head attention could make the model attend to the information from different representation subspaces of different tokens.

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\cdots,\text{head}_h)W^O,$$
$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (5)$$

where W_i^Q, W_i^H, W_i^V and W^O are learnable parameters. Given queries Q, keys K and values V, multi-head attention first projects Q, K and V for different heads with different linear projections W_i^Q, W_i^H, W_i^V. At each head, scaled dot-product attention is performed to obtain the attention output. Finally, the attention outputs of all heads are concatenated and then projected with W_i^O to the final outputs.

Model structure. As shown in Fig. 3, Transformer also consists of four basic components: the embedding layer, the encoder network, the decoder network, and the softmax layer.

1) Embedding layer: Similar to other NMT models, the embedding layer in Transformer also converts a source sentence $X = (x_0,\cdots,x_i,\cdots,x_m)$ into continuous embeddings $\boldsymbol{X} = (\boldsymbol{x}_0,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_m)$. After that, Transformer adds the positional embeddings to the input embeddings. They use sine and cosine functions to obtain the positional embeddings, which are then fed into the encoder layers.

2) Encoder network: The Transformer encoder consists of N identical layers. Each layer has two sub-layers. The first is a multi-head self-attention sub-layer, and the second is a feed-forward sub-layer. Then, residual connection and layer normalization are adopted to produce the final outputs. In the bottom multi-head self-attention sub-layer, the keys, values, and queries come from the positional embeddings. In the other multi-head self-attention sub-layers, all the keys, values, and queries come from the outputs of the previous layer of the encoder.

3) Decoder network: The decoder also consists of N identical layers. Different from the encoder layer, the decoder contains three sub-layers: The first is a masked multi-head self-attention sub-layer, and the second is a feed-forward sub-layer, and the third is encoder and decoder multi-head attention. Similar to the encoder, residual connections and layer normalization are also used to produce the outputs.

In the masked multi-head self-attention sub-layer, Transformer masks the subsequent embeddings to prevent the model from attending to subsequent tokens and ensures that at position j, the model can only utilize the information of the already produced outputs at positions less than j. In the encoder and decoder multi-head attention sub-layer, the queries are the output of previous decoder layer, and the memory keys and values come from the encoder. This allows the decoder to utilize the information from the source side.

4) Softmax layer: Similar to other NMT models, the outputs of the decoder are fed into a softmax layer to predict each token of the target sentence.

3 Main challenges and topics of NMT

NMT has made significant progress recently and has been widely utilized in many online MT systems[7, 8],

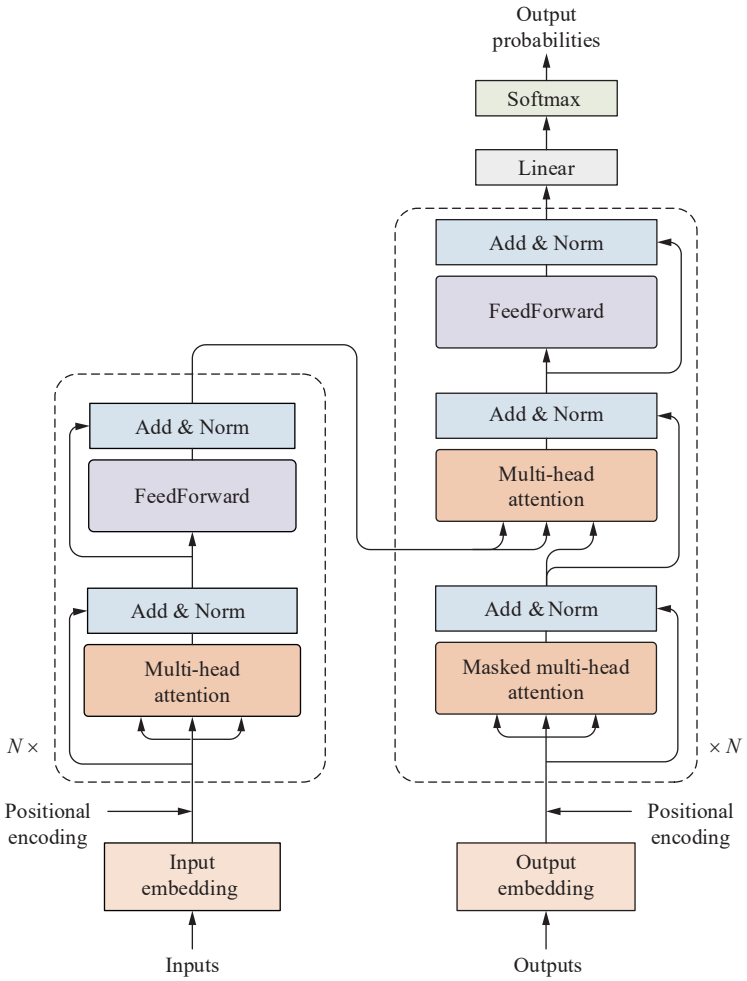


Fig. 3 Model structure of Transformer[10]

while the current NMT model still faces many challenges. In this section, we briefly introduce these challenges and topics of NMT.

Low-resource NMT. Current NMT models are data-driven methods and heavily rely on the quantity and quality of parallel training data. Only when the parallel training sentence pairs are frequent, can NMT learn the word and phrase translation pairs. Unfortunately, parallel training sentence pairs for NMT are expensive and hard to acquire in low-resource language pairs. Even in resource-rich language pairs, the training data are still inadequate in many domains, such as the medical domain, agricultural domain, and chemical domain. Therefore, improving the translation performance of low-resource NMT is a major challenge.

Document NMT. The current NMT usually takes sentences as the basic translation units. When fed an entire document, the current NMT models first split the document into sentences and then translate sentences in isolation without considering the cross-sentence dependencies. However, the absence of document context makes the model unable to deal with the document phenomena, such as deixis, lexical consistency, inflection, and omissions. Consequently, it is a major challenge for the NMT model to utilize document-level contextual information to improve translation quality over sentences in a document.

Multimodal NMT. Traditional NMT is a system whose input and output are both texts. However, human languages are not only about texts. Recently, there has been growing interest in multimodal machine translation, where the input contains other modalities, such as speech, image and video. Specifically, a typical MMT task is image-text translation, where the meaning of text may be ambiguous, and the help of images may be needed to determine the correct translation. Another major challenge is speech translation, which translates source language speech into target language text and has attracted much attention in recent years. The traditional speech translation system follows a pipeline framework, which contains an automatic speech recognition (ASR) module and a text machine translation module, leading to the problem of parameter redundancy, time delay, and error propagation. Thus it is a major challenge to build new neural architectures that can deal with multimodal inputs.

Beyond autoregressive decoding. The existing NMT model generates the target sentence token by token

from left to right, which is referred to left to right (L2R) autoregressive decoding. Although L2R autoregressive decoding is easy for the neural model to train and decode, it still contains the following two drawbacks: 1) Low parallelizability: It generates the i-th target token y_i only after all the previous target tokens ($y_{<i}$) have been predicted. 2) Limited context: The autoregressive manner predicts each output word only using previously generated outputs but cannot utilize the target-side future contexts $y_{>i}$. Thus it remains a challenge to build neural architectures that are beyond the current autoregressive decoding.

Prior knowledge integration. As we mentioned before, NMT is a data-driven method and it needs adequate parallel sentence pairs. In addition to the parallel sentence pairs, various prior knowledge (e.g., syntactic structure, bilingual lexicon and phrase, knowledge graphs) is also important for NMT. For example, incorporating an external bilingual lexicon could help the NMT model translate the low-frequency word. However, the current NMT network follows the encoder-decoder framework, which represents the tokens and semantics in distributed vectors. Prior knowledge is always represented by the discrete symbols, making it difficult to integrate prior knowledge into the current NMT framework. Accordingly, it remains a major challenge to integrate discrete symbol based knowledge into the distributed representation based NMT framework.

4 Representative approaches for each challenge

In this section, we mainly introduce the approaches for each challenge on the basis of Transformer. Meanwhile, we also represent the approaches on the basis of RNMT or ConS2S if these methods are very representative or model-agnostic.

4.1 Low-resource NMT

To improve the translation performance of low-resource NMT, there are several directions: 1) semi-supervised NMT, 2) unsupervised NMT, 3) multilingual NMT, and 4) pre-trained language models for NMT.

4.1.1 Semi-supervised NMT

Although parallel training data are difficult to acquire and expensive, monolingual training data are usually abundant and easy to obtain. As a result, it is important to boost the NMT models with monolingual data. Thus semi-supervised NMT methods have been proposed to incorporate the source and target monolingual data into NMT.

Incorporating target monolingual data. For target monolingual data, Sennrich et al.[13] proposed a back-translation (BT) method for NMT. They first trained a target-to-source NMT model using parallel training data and utilized this target-to-source NMT model to translate monolingual sentences and obtain synthetic parallel data. Then, they mixed the synthetic parallel data and original parallel data to learn the source-to-target NMT model. Due to its model agnostic and good performance, BT has been widely used, especially when only a small amount of parallel data is available[14]. Edunov et al.[15] investigated the BT method at a large scale, and their analysis results showed that sampling based data and noise beam search based synthetic data could produce better translation performance than argmax inference based synthetic data.

Incorporating source monolingual data. There are also some studies that explore source monolingual data to improve the translation quality of NMT. Zhang and Zong[16] proposed two different strategies to make full use of source monolingual data. The first is self-training, which is similar to the BT method and builds synthetic parallel data with a source-to-target translation model. The second is multi-task learning with a translation task and a source-side reordering task.

Incorporating both source and target monolingual data. Many researchers have tried to make full use of both source- and target-side monolingual data in NMT. Cheng et al.[17] presented a semi-supervised approach for NMT, whose main idea is to reconstruct the source and target monolingual corpora using an auto-encoder. He et al.[18] proposed a dual-learning approach for NMT. Their idea is inspired by the observation that machine translation tasks have a dual task, i.e., a primal task from source to target translation and a dual task from target to source translation. These two tasks could form a closed loop and help each other. Thus they propose a dual-learning algorithm to teach each other through a reinforcement learning process.

4.1.2 Unsupervised NMT

Unsupervised neural machine translation (UNMT)[19, 20] considers a more challenging scenario in which parallel sentence pairs are unavailable, and there are only massive source-side monolingual data and target-side monolingual data.

Early studies focused on the bilingual lexicon induction (BLI) task, which is a word-level translation task in unsupervised scenarios. BLI aims at inducing word translations with only monolingual corpora of two languages. At present BLI has become an important component for UNMT. Mikolov et al.[21] proposed a method to extend bilingual dictionaries based on large monolingual data. It learns a linear mapping to transform the source embeddings to target embeddings by minimizing the distance between the bilingual seeds. Subsequent studies totally eliminate the bilingual seed dictionary and learn the mapping function in a purely unsupervised way[22–24].

Motivated by BLI, various UNMT methods[19, 20] have been proposed to achieve the sentence-level translation. Artetxe et al.[19] proposed a UNMT with denoising and back-translation. Fig. 4 shows its architecture. For a sen-

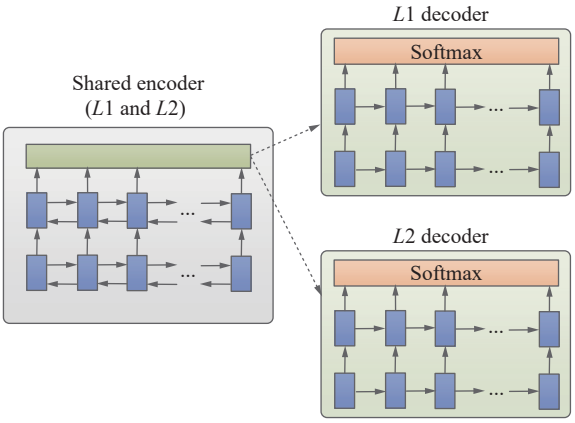


Fig. 4 Model structure of the UNMT in [19] with two steps: The denoising step reconstructs the sentence x from its noised version. The back-translation step translates the sentence x_{L1} to another language x_{L2}, and then inputs x_{L2} with the shared encoder and decoder for $L1$ to recover the original sentence x_{L1}.

tence x_{L1} in language $L1$, the proposed method contains two-steps to train the model:

1) Denoising step, which reconstructs the sentence from its noised version x'_{L1} with a shared encoder and decoder of language $L1$. The produce can be illustrated as follows:

$$x'_{L1} \rightarrow \text{Enc}_{\text{share}} \rightarrow \text{Dec}_{L1} \rightarrow x_{L1} \tag{6}$$

where $\text{Enc}_{\text{share}}$ is the shared encoder for $L1$ and $L2$. Dec_{L1} is the decoder for $L1$.

2) Back-translation step, which translates the sentence x_{L1} to another language x_{L2}, and then input x_{L2} with the shared encoder and decoder for $L1$ to recover the original sentence x_{L1} as follows:

$$\begin{aligned} &x_{L1} \rightarrow \text{Enc}_{\text{share}} \rightarrow \text{Dec}_{L2} \rightarrow x_{L2} \rightarrow \\ &\text{Enc}_{\text{share}} \rightarrow \text{Dec}_{L1} \rightarrow x_{L1}. \end{aligned} \tag{7}$$

For a sentence x_{L2} in language $L2$, the same two steps are conducted. By iterating above two-steps for $L1$ and $L2$, a UNMT model could be learned with only monolingual data.

In parallel, Lample et al.[20] also proposed a UNMT model, where the model starts with an unsupervised initial translation model via a word-by-word translation learned in an unsupervised way. Then, the model is trained by a reconstruction task and a back-translation task. Meanwhile, they also proposed a discriminator to improve the alignment of sentences in the source and the target languages.

Furthermore, Artetxe et al.[25] and Lample et al.[26] focused on the traditional SMT and proposed an unsupervised statistical machine translation (USMT), whose performance is comparable with that of UNMT. Since USMT and UNMT were proposed, several studies[27, 28] have tried to combine UNMT and USMT to improve unsupervised machine translation performance.

Recently, various methods have addressed the multilingual UNMT scenario, where there are some language pairs for auxiliary languages and monolingual for unsupervised languages, its goal is to learn the translation model for unsupervised languages with auxiliary parallel data[29–32].

4.1.3 Multilingual NMT

Standard NMT can only translate a source sentence into another target sentence. Although achieving promising results, it is inconvenient to train each separate NMT model for each language pair, especially when there is a demand to translate between hundreds of languages. Multilingual neural machine translation (MNMT)[33, 34] aim to build a unified NMT model to translate multiple languages. MNMT can not only improve the translation performance through knowledge transfer, but also facilitate model deployment. Fig. 5 shows the comparison of NMT (Fig. 5(a)), a completely shared MNMT[34] (Fig. 5(b)) and MNMT with language-independent and language-specific parameters (Fig. 5(c)).

Johnson et al.[34] proposed a simple but effective universal MNMT method. In this method, there is no need to change the network architecture. The only modification is that they introduce a special indicator at the beginning of the source sentence to indicate source and target language. For example, consider the following English-to-Italy sentence pair:

you probably saw it on the news. → forse lo avete visto sui notiziari.

It will be modified to:

$\langle 2it \rangle$ you probably saw it on the news. → forse lo avete visto sui notiziari.

Where $\langle 2it \rangle$ is an indicator to show that the target is Italy. After adding the tokens to multilingual training data, they train the MNMT model with all multilingual language pairs, where all source languages share the same encoder and all the target languages share the same decoder.

Due to its simplicity and low-resource language improvement, the universal MNMT has drawn much attention. Massively multilingual experimental results show that the completely shared model faces capacity bottlenecks for retaining the translation performance of each language[35]. Thus, various researchers have tried to balance the language-independent and the language-specific parameters in a whole model[36–38]. Bapna and Firat[38] proposed an adaptation approach for MNMT, which injects language specific adapter layers into a pre-trained MNMT model. These adapters could adapt the model to multiple individual language pairs simultaneously. Eriguchi et al.[39] proposed a two-stage training for MNMT that serves an arbitrary task-specific translation direction, which first pre-trains an MNMT model and then fine-tunes the model to the task-specific MNMT model.

Instead of designing language-specific parameters manually, some subsequent studies have attempted to

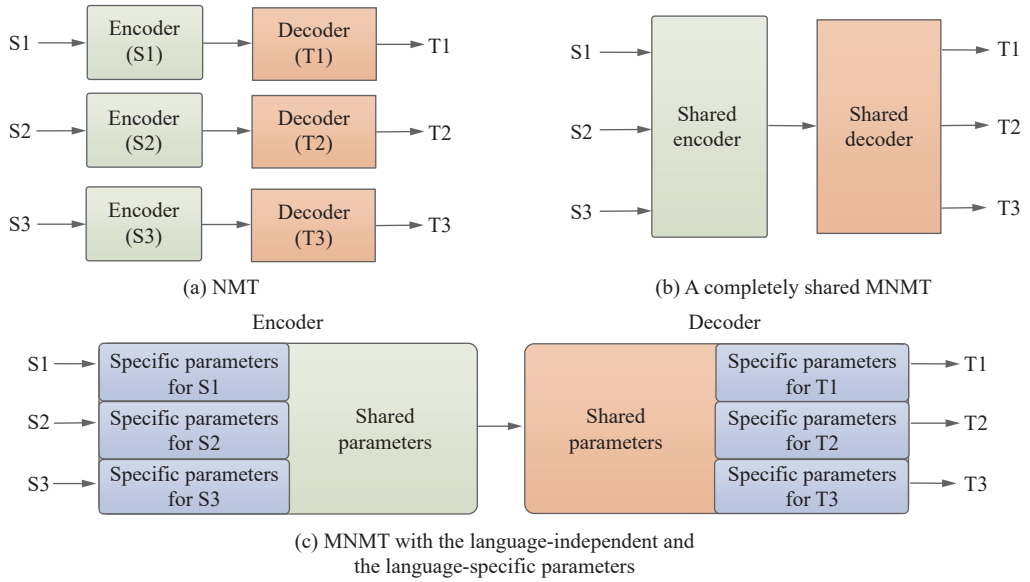


Fig. 5 Illustration of MNMT, where S1, S2 and S3 denote the three different source languages. Similarly, T1, T2 and T3 denote three different target languages. In standard NMT (a), we need to train each separate neural model for each language pair. In universal MNMT (b), all parameters are shared by all language pairs. In MNMT with language-independent and language-specific parameters (c), all parameters are divided into two parts: i) language-independent parameters to model the shared knowledge of all languages; ii) language-specific parameters to model the specific knowledge for each language.

search for a language-specific subspace of the whole model[40–42]. Lin et al.[41] proposed a method that dynamically learns language specific sub-network (LaSS) for MNMT, in which each sub-network shares partial parameters with some other languages while also retaining its language-specific parameters. Wang and Zhang[42] proposed a parameter differentiation based MNMT to make the model decide which parameters should be language-specific and which ones should be shared. Their model is initially completely shared by all languages, and then the model detects shared parameters that should be language-specific.

4.1.4 Pre-trained language models for NMT

Pre-trained language models (PTMs) are alternative ways to improve the translation performance of low-resource NMT. Recently, pre-training techniques have attracted much attention in natural language processing communities. The pre-trained models first learn the universal language representations through various pre-training tasks, and then fine-tuning methods are utilized to transfer the knowledge in pre-trained models to the downstream tasks. Qiu et al.[43] conducted a comprehensive overview of pre-trained language model for natural language processing.

Most of the early PTMs, such as ELMo[44] and BERT[45], achieve state-of-the-art performance in various language understanding tasks such as sentiment classification, natural language inference, and named entity recognition. Inspired by their success, many methods introduce these PTMs to NMT. Edunov et al.[46] incorporated the ELMo[44] into the NMT model by 1) inputting ELMo embeddings into the encoder and the decoder, and 2) finetuning the ELMo parameters with parallel sentence pairs. After that, the researchers attempted to enhance the NMT with BERT. Zhu et al.[47] proposed a BERT-fused translation model, in which it first utilizes the BERT to extract representations for a source sequence. Then, the representations are fused with each layer of the encoder and decoder in NMT via attention mechanisms. Similarly, Yang et al.[48] proposed three techniques (asymptotic distillation, dynamic switching, and rate-scheduled learning) to integrate BERT and NMT.

Considering that directly applying these BERT like pre-training methods on the natural language generation tasks, including machine translation tasks, is still inconvenient, thus some sequence-to-sequence pre-training methods have been proposed, such as MASS[49], T5[50] and BART[51].

MASS[49] (Fig. 6) adopts the encoder-decoder framework, where the encoder takes a sentence with a masked fragment as an input, and the decoder recovers the masked fragment. MASS can be utilized in various natural language generation tasks, such as NMT and text summarization. BART[51], a pre-trained model combining bidirectional and auto-regressive Transformers, is a denoising auto-encoder with sequence-to-sequence models. Specifically, BART is pre-trained in two stages: 1) corrupting the original text with a noise function, and 2) learning a reconstruction model to reconstruct the original text. Wang et al.[52] studied the impact of the jointly pre-trained decoder, and proposed an in-domain pre-training and input adaptation strategy to overcome the domain and objective discrepancies.

Different from the above pre-trained language models in a single language, the recent work attempts to build many cross-lingual or multilingual pre-trained models to make the model learn the cross-lingual or multilingual

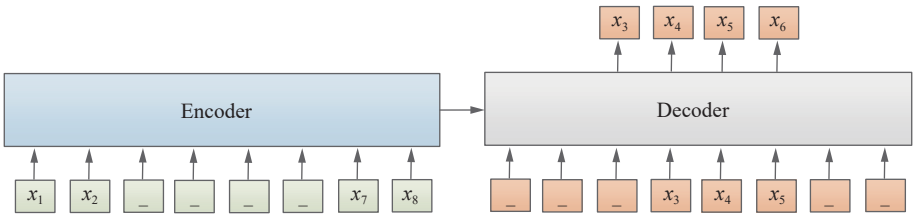


Fig. 6 Framework for MASS[49], which adopts the encoder-decoder framework, where the encoder takes a sentence with a masked fragment as input, and the decoder recovers the masked fragment. The token "_" denotes the mask symbol.

knowledge during the pre-training phase, such as mBERT[53], XLMs[54], mRASP[55], mT6[56], mBART[57] and CeMAT[58]. Take the mBART as an example. mBART[57] is a multilingual sequence-to-sequence denoising auto-encoder pre-training model, which extends BART[51] to many languages. The encoder takes the input texts with various noises, and the decoder recovers the original text. mBART is first pre-trained once for all available languages, then it can be fine-tuned by parallel sentence pairs for the NMT tasks.

4.2 Document NMT

Standard NMT methods usually focus on sentence-level translation. However, this sentence-level translation cannot address document-level translation. To utilize the cross-sentence context, many researchers propose document-level neural machine translation (DocNMT), which could improve the translation quality with the help of the context information in the document. As shown in Fig. 7, the current DocNMT can be roughly divided into two categories: DocNMT with source-side context[59–62] and DocNMT with target-side context[63–65].

DocNMT with source-side context. As the name suggests, given a source sentence, DocNMT with source-side context methods could utilize more source-side context to improve the translation quality. These methods can be divided into two kinds: 1) single encoder models and 2) multi-encoder models.

The representative single encoder model was proposed by Bawden et al.[59], where they first concatenated a source sentence with its source-side context and then input this long sequence into the neural model. Based on a single encoder model, Zhang et al.[62] studied the transfer of contextual information via multilingual transfer from document-rich languages to document-poor languages.

Different from the single encoder model, multi-encoder model[60, 61] encodes a source sentence and its contexts with different encoders. Voita et al.[60] proposed a context-aware DocNMT on the basis of Transformer. In the model, a source sentence and its contexts are first encoded independently. Then, an attention layer is used to obtain a context-aware semantic representation, which is utilized to produce the target sentence. Zhang et al.[61] extend the Transformer with a new context encoder and proposed a two-step training strategy to fully utilize document-level and sentence-level parallel training data. Lupo et al.[66] proposed a divide and rule pre-training approach to enhance the contextual training signal.

DocNMT with target-side context. In addition to the source-side context, researchers also utilize the target-side context to improve the DocNMT. These methods contain two different kinds of methods: 1) the context-cache model[63, 67] and 2) the two-pass decoder model[64, 65].

The context-cache model[63, 67] utilizes a cache to store the translation history, which is then used to improve the

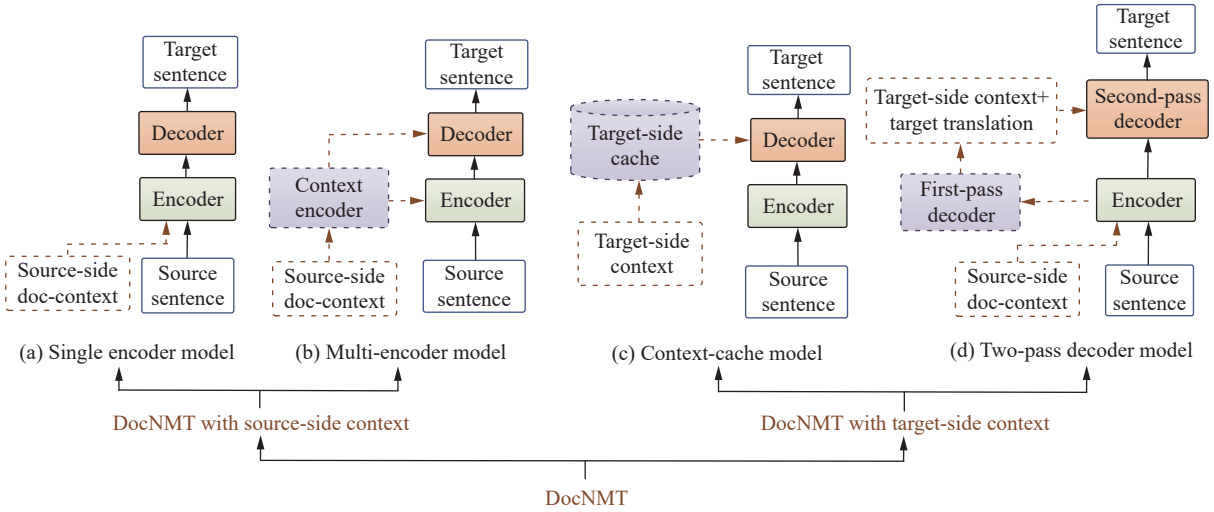


Fig. 7 Two categories of DocNMT: 1) DocNMT with source-side context and 2) DocNMT with target-side context.

translation performance. For example, Tu et al.[63] proposed an approach to augment NMT with a cache-like memory, which stores the target-side translation history.

Another line of DocNMT with target-side context is the two-pass decoder model. Voita et al.[64] introduced a two-pass DocNMT framework, where a source sentence is first translated with a context-agnostic model (first-pass), and the translation is then refined with the context of several previous sentences (second-pass). Furthermore, Voita et al.[65] proposed a repair approach for DocNMT, which could reduce the inconsistencies between sentence-level translations with only monolingual document-level data. Meanwhile, dynamic context methods[68] are proposed to select the useful document context for different DocNMT models.

4.3 Multimodal machine translation

The current NMT models mainly focus on the text translation scenario where both input and output are text sentences. However, in practice, there has been growing interest in multimodal machine translation (MMT), where the input contains other modalities, such as speech, images and videos. Here we mainly introduce the two tasks in MMT: 1) image-text translation and 2) speech translation.

4.3.1 Image-text translation

The setting of image-text translation (also named as image caption translation) is that the inputs of the translation model contain two modals: an image and its text description in source language. It needs to translate the description in the source language with the help of an image into the target language. Depending on the utilization of visual information, the current image-text translation methods can be categorized as 1) coarse-grained utilization and 2) fine-grained utilization.

Coarse-grained utilization methods[69–71] represent the image in the global semantics and then incorporate it by inputting it as an auxiliary or attending to relevant local regions. Huang et al.[69] enhanced the attention-based RNMT model by incorporating information in multiple modalities (image and text). They transform the visual features as one of the steps in the encoder, and then attend to both the text and the image during decoding. Calixto et al.[70] proposed a latent variable model to model the visual and textual features for image-text translation. Their model could utilize the visual and textual inputs during training but does not require images during testing. Ive et al.[71] proposed a translate-and-refine approach, where they generate the first draft only by the source text, and then improve the first draft with the target language and visual context.

Fine-grained utilization methods[72–74] aim at incorporating the entity-level or object-level visual context to improve the translation. Huang et al.[72] proposed an explicit entity-level cross-modal learning method to augment the entity representation, where a multi-task method is utilized to enhance entity representation and improve text translation. Wang and Xiong[73] presented an object-level visual context framework, that models the interaction between the visual and text modality by using the object-text similarity and object-source attention. Li et al.[74] proposed a selective attention multimodal Transformer and utilize vision Transformer to extract vision features.

4.3.2 Speech translation

Speech translation aim to translate a speech in the source language into text in the target language, which could facilitate the communication between two speakers of different languages. The traditional speech translation system follows a pipeline framework, which is composed of an ASR module and a text MT module. In the pipeline system, two modules are learned independently, leading to the problem of time delay and parameter redundancy. To overcome these problems, end-to-end speech translation[75–78], which directly translates from source language speech to target language texts, has attracted much attention in recent years. Thus here we focus on the end-to-end speech translation methods.

Generally, the current speech translation methods can be divided into the following two types: 1) incorporating the data of ASR and text translation, and 2) bringing the feature gap.

Incorporating the data of ASR and NMT. To train an end-to-end speech translation model, we need parallel data with audio signals and their corresponding target translations. Considering the low resources of training data for speech translation, many researchers have tried to incorporate the training data of ASR and NMT into speech translation tasks to improve the performance. Multi-task learning and data augmentation[75, 76, 79], pre-training[80–83] and knowledge distillation[84–87] are the three main directions.

Multi-task learning for speech translation could better optimize the parameters of end-to-end model with the auxiliary tasks, i.e., ASR tasks or NMT tasks[75, 76]. Data augmentation[79] tends to generate synthetic data for speech translation.

In the pre-training framework, the parameters are first pre-trained by the ASR training data or MT training data, and then the pre-trained parameters are utilized as a parameter initialization of the speech translation model[80–83]. Bansal et al.[80] proposed a pre-training method with ASR training data, which first pre-trains the parameters on a high-resource ASR task, and then fine-tunes the parameters with speech translation dataset. Wang et al.[81] proposed a tandem connectionist encoding network to reuse all subnets in pre-training, keep the roles of subnets in consistent, and pre-train the attention module. WavLM in [82] first learns universal speech representations using large-scale unlabelled speech data, and then adapts the representations to various speech processing tasks. Tang et al.[83] proposed a speech and text

joint pre-training to combine the pre-training method and multi-task method.

The third direction is knowledge distillation based methods[84–87]. Liu et al.[84] proposed a knowledge distillation based speech translation, where the speech translation is a student model, whose input is a speech. The machine translation model is a teacher model, whose input is a source text. By utilizing a knowledge distillation method, the speech translation model learns by not only the correct target texts, but also the output probabilities produced by the machine translation model. Inaguma et al.[85] proposed a bidirectional knowledge distillation method, in which they combine a source-to-target NMT model and a target-to-source NMT model to distill the knowledge from both forward and backward NMT models to the end-to-end speech translation model. Ren et al.[87] introduced two different knowledge distillation methods for speech translation: attention-level knowledge distillation and data-level knowledge distillation.

Bringing the feature gap. The traditional speech translation model utilizes the frame-level features as speech representations. The frame-level features create longer, sparser input sequences than their text equivalents[77, 88, 89]. This increases the memory usage and model parameters, and thereby impacts the translation quality. Accordingly, some studies have been proposed to bridge the gap between frame-level speech features and word-level text features. Salesky et al.[88] explored the phoneme-level speech representations for speech translation. They first utilized the alignment methods to generate phoneme labels, and then created phoneme-level feature representations from variable numbers of frames. Ye et al.[77] designed a contrastive learning method for speech translation to reduce the gap between the representations of speech and the corresponding transcription.

4.4 Beyond autoregressive decoding

As we introduce in Section 3, the existing NMT model generates the target sentence token by token from left-to-right in an autoregressive manner, which contains two drawbacks, i.e., low parallelizability and limited context. In recent years, many methods beyond autoregressive decoding have been proposed to overcome the drawbacks of autoregressive decoding. Generally, non-autoregressive Transformer (NAT)[90] and bidirectional decoding[91–93] are two main research lines, where NAT can generate all target words in parallel, and bidirectional decoding can simultaneously predict the target sentences from both left-to-right and right-to-left directions.

4.4.1 Non-autoregressive decoding

To speed up the generation and improve the parallelizability, Gu et al.[90] first proposed a NAT model to decode target tokens in parallel. Given a source sentence x and target sentence y, NAT calculates the probability $p(y \mid x; \theta)$ as follows:

$$p(y \mid x; \theta) = p_l(T \mid x; \theta) \prod_{i=1}^{T} p(y_i \mid x; \theta) \tag{8}$$

where $p_l(T \mid x; \theta)$ determines the length of target sentence T given x. To achieve this, NAT proposes a fertility predictor, where it first predicts the fertility $\Theta(x_j)$ (the number of target words that should be translated) for each source word x_j. Then the total length is calculated by $T = \sum \Theta(x_j)$. Finally, NAT copies the source tokens based on predicted fertilities and generates the target sentence in parallel. Meanwhile Gu et al.[90] also proposed a knowledge distillation method to transfer the knowledge from autoregressive Transformer (AT) models to NAT models. The idea of knowledge distillation has been widely adopted in many studies[94, 95]. Ding et al.[96] proposed a monolingual knowledge distillation model, that leverages monolingual data to perform knowledge distillation, and trains the NAT model with distilled monolingual data. Shao et al.[97] proposed a diverse distillation with reference selection for NAT.

Although the remarkable efficiency has been achieved by NAT, it still suffers from quality degradation compared to the AT models. Thus, many methods have been proposed to improve the translation of NAT[98–101]. Wang et al.[98] proposed a semi-autoregressive Transformer-model (SAT), which keeps the AT global while conducts NAT locally. Gu et al.[100] proposed a Levenshtein Transformer, which is a partially autoregressive model with the insertion and deletion operations.

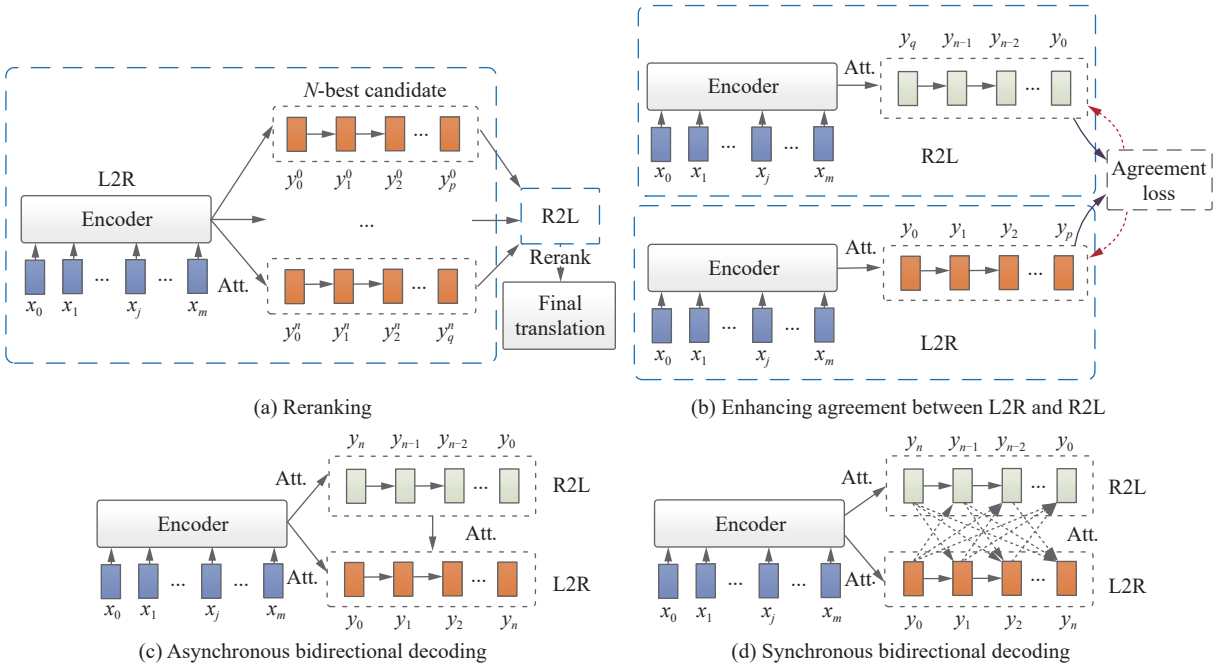
There are also some studies to improve NAT with multi-pass iterative refinement. Lee et al.[102] proposed a conditional NAT with iterative refinement, which is designed based on latent variable models and denoising autoencoders. Zeng et al.[103] proposed a plugin algorithm, called as aligned constrained training, to handle low-frequency constraints in NAT.

4.4.2 Bidirectional decoding

Some methods are proposed to combine the merits of L2R decoding and R2L decoding. As shown in Fig. 8, these methods can be divided into four categories: 1) reranking, 2) enhancing agreement between L2R and R2L, 3) asynchronous bidirectional decoding, and 4) synchronous bidirectional decoding.

The first common studies to combine L2R and R2L models are reranking[104, 105]. Reranking techniques have been widely applied in SMT systems. In regard to NMT, the L2R model first generates n-best target candidates. Then, the R2L score is utilized to select the best candidate as the final translation.

Another branch of bidirectional decoding is enhancing agreement between L2R and R2L[8, 106, 107]. Intuitively, the translation results from L2R and R2L decoding should be the same. Thus, agreement information is utilized to encourage the neural model to produce better translations.



(a) Reranking (b) Enhancing agreement between L2R and R2L

(c) Asynchronous bidirectional decoding (d) Synchronous bidirectional decoding

Fig. 8 Four different kinds of methods to achieve bidirectional decoding

Subsequent studies[91, 108] proposed an asynchronous bidirectional decoding to combine the L2R and R2L models. Specifically, Zhang et al.[91] added an R2L (or backwards) decoder into the standard NMT model. Fig. 8(c) shows the framework, which mainly contains three components: 1) an encoder, which encodes the source sentence into the hidden states; 2) an R2L decoder, which translates target translation in the R2L manner; and 3) an L2R decoder, which produces the final translation using L2R decoding with source-side hidden states and R2L hidden states.

Different from asynchronous bidirectional decoding, Zhou et al.[92] proposed a synchronous bidirectional decoding for NMT, which could predict final outputs using L2R and R2L decoding simultaneously and interactively. Fig. 8(d) shows the framework, where the L2R decoding could generate each token y_i not only depends on previously generated tokens (i.e. $y_1, y_2, \cdots, y_{i-1}$), but also relies on previously predetermined target tokens (i.e., y_n, $y_{n-1}, \cdots, y_{n-i+1}$) of R2L decoding. Meanwhile, R2L decoding also generates each target token in the same way. Furthermore, Zhou et al.[109] extended synchronous bidirectional decoding by producing target sentences from both sides (L2R and R2L) to the middle, which significantly speeds up decoding time and also improves the generation quality.

4.5 Prior knowledge integration

In addition to parallel sentence pairs, prior knowledge is also an important resource for NMT. According to the different types of knowledge, these studies can be divided into three types: 1) bilingual lexicons, 2) phrase tables and terminologies, and 3) knowledge graphs.

Bilingual lexicons in NMT. Some experimental results[110, 111] show that the current NMT model has problems in handling low-frequency words. To improve the translation accuracy of low-frequency words, many researchers tend to incorporate bilingual lexicons into NMT[110–113][1]. Zhang et al.[112] proposed to a posterior regularization method to integrate prior knowledge into NMT. They represented the prior knowledge, including bilingual dictionary, phrase table, coverage penalty and length ratio, in a log-linear model to guide the learning of NMT model. Arthur et al.[110] focused on low-frequency words and proposed a method to incorporate discrete bilingual lexicons into NMT. Feng et al.[111] proposed a memory-enhanced NMT, where each memory element stores a source-target pair, and memory is utilized by the attention layer to improve the neural model.

Phrase table and terminologies in NMT. The current NMT generates the target sentence word-by-word, making it difficult for the neural model to translate phrases or terminologies with multi-words. To better deal with phrases or terminologies, some researchers tend to utilize the phrase table[116–119] and terminologies[120, 121] in NMT. Wang et al.[116] proposed a method to integrate a phrase memory from the SMT model into the NMT model. Specifically, the SMT model dynamically generates phrase memory. Then, the decoder selects a phrase from the

[1] Another solution to overcome the problem of low-frequency and out-of-vocabulary words is to decompose these words into smaller granularities, such as hybrid word-character[114], sub-word[115] or word piece[7]. At present, sub-word methods[115] have been widely used to address the low-frequency words and open vocabulary problem.

memory or a word from the neural model with a balancer. Dahlmann et al.[117] proposed a hybrid search that extends NMT beam search with phrase-based models. During the beam search, their method inserts phrasal translations into NMT, and then the various scores from the SMT and NMT are utilized in a log-linear model to decide the final translation. Xu et al.[119] proposed attentive phrase representations for NMT, where it first generates phrase representations from the corresponding token representations, and then incorporates the phrase representations into Transformer to capture long-distance relationships.

Knowledge graphs in NMT. Knowledge graphs (KGs) contains lots of structured facts on entities. The current KG is always organized with triples (h, r, t), where h and t indicate head and tail entities, and r denotes their relation. To utilize these KGs and enhance the entity translation in KGs, many studies incorporate KGs into NMT[122–124]. Zhao et al.[122] proposed a KG enhanced NMT method. It first induces the translation results of new entities via entity alignment. Then, it generates pseudo parallel sentence pairs with the induced entity pairs. Finally, the NMT model is jointly learned with the original and pseudo parallel sentence pairs. DEEP[124] is a denoising entity pretraining method to leverage monolingual data and a knowledge base to improve named entity translation accuracy. Meanwhile, to further improve the translation quality of entities, it also utilizes the multi-task learning to fine-tune a pre-trained NMT model on both entity-augmented monolingual data and parallel data.

Table 1 summarizes the mentioned challenges in NMT and the corresponding partial representative methods. There are still many topics which are involved in NMT while not mentioned in this article, such as robustness, interpretability, and enhancing the current training objectives.

5 Resources and tools

Thanks to the open research environment in the NMT community, there are many open resources and toolkits, which significantly improve development. Table 2 lists the available and useful corpora for different machine translation tasks.

Text parallel data. There are several widely used text parallel data for machine translation tasks. The Conference on Machine Translation (WMT)[2] conducts different shared translation tasks and releases the corresponding parallel data in each year. In addition, the *International Conference on Spoken Language Translation* (IWSLT)[3] is also an annual scientific conference, that organized many evaluation campaigns and published a large amount of parallel data. The *China Conference on Machine Translation* (CCMT)[4], organized by the Chinese Information Processing Society of China, also conducts machine translation evaluations each year. The evaluations in CCMT mainly focus on Chinese-centric translation tasks (such as Chinese-to-English, Mongolian-to-Chinese, Tibetan-to-Chinese, etc.) in different domains (such as daily conversation translation, government document translation, news translation, etc.). All evaluation participants in CCMT could acquire the corresponding parallel data for free.

We can also find the parallel data from the LDC[5] and Europarl[6]. Meanwhile, a widely used project is OPUS[7], which is a growing collection of parallel texts from the web.

Image-text translation data. Compared to text parallel data, the image-text translation data are quite rare. The most widely used is the Multi30k dataset[127], which extends the Flickr30K dataset (the original descriptions are English) with German translations and German descriptions.

Speech translation data. The widely used end-to-end datasets for speech translation are Fisher and Callhome[131], TED[84][8], Augmented LibriSpeech[128] and MuST-C[129].

Document translation data. The parallel document-level data mainly include TED[130], News-Commentary[9] and Europarl[126].

There are also many open-source toolkits for the research and development of NMT systems. Table 3 provides some broadly used toolkits, including the programming language and the framework. In addition, there are some other survey papers on NMT[132–134].

6 Extensions to other tasks

Although Transformer was originally proposed for NMT, it now goes far beyond MT and extends to more tasks. Many studies apply Transformer to different tasks and achieve state-of-the-art performance. In this section, we briefly introduce the application of Transformer in other tasks, including NLP tasks (such as pre-trained language models, summarization, knowledge graphs, dialogue), the computer vision tasks (image classification, image generation), audio tasks (automatic speech recognition and speech synthesis) and multi-modal tasks.

[2] WMT was held 10 times as a workshop from 2006 to 2015. Since 2016, it is was held as a conference in Berlin, Germany. The latest WMT 2022 can be found in https://www.statmt.org/wmt22/index.html.

[3] https://iwslt.org/

[4] The latest CCMT 2022 can be found in http://sc.cipsc.org.cn/mt/conference/2022/

[5] https://www.ldc.upenn.edu/

[6] https://www.statmt.org/europarl/

[7] https://opus.nlpl.eu/

[8] http://www.nlpr.ia.ac.cn/cip/dataset.htm.

[9] https://www.casmacat.eu/corpus/news-commentary.html

Table 1 Challenges in NMT and partial representative methods

Challenges	Main directions	Description	Methods
Low-resource NMT	Semi-supervised NMT	Improving NMT with source and target monolingual data	Using source data[16]
			Using target data[13]
			Using both source and target data[17, 18]
	Unsupervised NMT	Building a NMT model with only massive source and target monolingual data	Unsupervised NMT[19, 20]
			Unsupervised SMT[25, 26]
			Combining UNMT and USMT[27, 28]
			Multilingual UNMT[29, 30]
	Multilingual NMT	Building a unified NMT model to translate multiple languages	Parameters completely shared[34]
			Balance the language-independent and language-specific model[40–42]
	Pre-training for NMT	Enhancing NMT with pre-trained language models	Utilizing ELMo[46] and BERT[47, 48]
			Sequence-to-sequence pre-training[49, 51]
			Multilingual and cross-lingual pre-training[53–55]
Document NMT	DocNMT	Enhancing NMT with document context	DocNMT with source-side context[59–62]
			DocNMT with target-side context[63, 64]
Multi-modal NMT	Image-text NMT	Translating a source sentence and an image into the target	Coarse-grained utilization[69–71]
			Fine-grained utilization[72–74]
	Speech translation	Translating a speech in source language into a text in target language	Incorporating the data of ASR and NMT[80–85]
			Bringing the feature gap[77, 88, 89]
Beyond autoregressive decoding	Non-autoregressive decoding	Generating all target words in parallel	NAT[90]
			Semi-NAT[98]
			NAT with multi-pass iterative refinement[99, 102]
	Bidirectional decoding	Combining the left-to-right and right-to-left decoding	Reranking[104, 105]
			Enhancing agreement[8, 106, 107]
			Asynchronous bidirectional decoding[91]
			Synchronous bidirectional decoding[92]
Prior knowledge integration	Knowledge enhanced NMT	Enhancing NMT with various symbolic knowledge	Bilingual lexicons[110, 111]
			Phrase tables[116–119] and terminologies[120, 121]
			Knowledge graphs[122-124]

6.1 Extensions to other NLP tasks

Transformer for pre-trained language models. A recent huge breakthrough in NLP is pre-trained language models, which are first pre-trained on a large corpus and then utilized in different downstream NLP tasks to avoid learning a model from scratch[135]. Currently, Transformer has become the mainstream architecture for pre-trained language models in three different ways:

1) Transformer encoder: Many pre-trained language models for the natural language understanding tasks use the Transformer encoder as their architectures. The most repressive model is BERT[45], which utilizes masked language modelling and next sentence prediction as the pre-training objective. Furthermore, many BERT-like models have been proposed on the basis of BERT, such as RoBERTa[136] and SpanBERT[137].

2) Transformer decoder: Meanwhile, some studies utilize the Transformer decoder as the backbone architectures. For example, the generative pre-trained Transformer (GPT)[138] and its subsequent versions (GPT-2[139] and GPT-3[140]) pre-trained the language models in a zero-shot setting with the Transformer decoder.

3) Transformer encoder-decoder: There are also some pre-trained language models with the standard Transformer encoder-decoder architecture, such as MASS[49], T5[50] and BART[51], which have been introduced in Section 4.1.4.

Transformer for summarization. Text summarization is an important NLP task to generate natural lan-

Table 2 Some available resources for NMT

Types	Corpus name	Main language pairs
Text parallel data	WMT	*
	IWSLT	*
	CCMT	*
	LDC	*
	Europarl[125]	*
	OPUS[126]	*
Image-text translation data	Multi30K[127]	En ⇒ De
Speech translation data	Augmented LibriSpeech[128]	
	TED[84]	En ⇒ De/Fr/Zh/Ja
	MuST-C[129]	En ⇒ 14 languages[10]
	Fisher and callhome	En ⇒ Es
Document translation data	TED[130]	*
	News-commentary	*
	Europarl[125]	*

guage summaries that compress the information in a longer text. The current summarization also follows the encoder-decoder framework. Different NMT, the encoder-decoder framework in text summarization encodes a document and then generates summary with the decoder. Inspired by the remarkable progress of Transformers in NMT, many methods also apply the Transformers to text summarization tasks. Liu et al.[141] considered Wikipedia as a multi-document summarization where the inputs are topics and documents, and the target is the Wikipedia text. They utilize Transformer decoder to achieve this goal. Recently, many approaches have applied the Transformer-based pre-trained language model to summarization, such as BART[51] and PRIMERA[142]. PRIMERA[142] uses the gap sentence generation objective with an entity pyramid, in which the model masks salient sentences then it generates these important sentences.

Transformer for knowledge graph learning. Knowledge graphs contain much structured human knowledge and have drawn great research attention in NLP and knowledge mining[143]. Transformer is also widely utilized in recent advances of knowledge graph research. Here, we introduce the application of Transformer in knowledge representation learning. Knowledge representation learning (KRL) aims at mapping entities and relations into low-dimensional vectors while capturing their semantic meanings. The Transformer in KRL is helpful in compositional representation models or contextualized representation model. In compositional representation models, entity embeddings are composed of token embeddings. Saxena et al.[144] treated knowledge link prediction as sequence-to-sequence tasks and then train a Transformer to learn the knowledge representation and compositional functions. Other models utilize the Transformer to get the contextualized knowledge representation. CoLAKE[145] dynamically represents an entity with the knowledge context and language context and jointly learns knowledge and language representation in a unified representation space.

Transformer for dialogue. Dialogue systems are a popular topic in NLP and are widely used in daily life. Dialogue systems contain two kinds of research lines: task-oriented dialogue systems, which serve as assistants to solve specific tasks, and open-domain dialogue systems, which communicate with human without task restrictions. Transformers are widely utilized in both task-oriented dialogue systems and open-domain dialogue systems. Henderson et al.[146] proposed a response selection model for task-oriented dialogue with the Transformer framework, which first pretrains the response selection model with general-domain corpora and then fine-tunes the pretrained model with the target dialogue domain. Zhang et al.[147] proposed a dialogue generative pre-trained Transformer, which is trained on large-scale dialogue pairs and sessions from Reddit. Ma et al.[148] focused on the multimodal task-oriented dialog systems and propose a unified Transformer semantic representation framework, where a Transformer encoder is used to project all the multimodal features into a unified semantic space and a hierarchical Transformer response decoder is utilized to generate target text responses.

Transformer for question answering. Textual question answering (QA) aim to provide precise answers to questions from users in natural language. Specifically, textual QA is studied under two task settings based on contextual information, i.e., machine reading comprehension (MRC) and open-domain QA (OpenQA). Specifically, MRC tends to make machines answer a question given specified context passages. In contrast, OpenQA tries to answer a given question without any specified context. It requires the model to first search for the relevant documents as the context. Bao et al.[149] analysed matching components and Transformer encoding for MRC. Recently, Transformer-based pre-trained languages models have also been widely utilized in MRC and have achieved remarkable progress[150]. For OpenQA task, instead of extracting answer spans, researchers utilize Transformer to generate the answers. Masque[151] utilizes the Transformer and multi-source abstractive summarization to learn the multi-style answers for OpenQA. Transformer-XH[152] boosts Transformer with structured text data for complex reasoning OpenQA. Similar to the other tasks, some OpenQA systems[153] also borrow Trans-

[10] Arabic, Chinese, Czech, Dutch, French, German, Italian, Persian, Portuguese, Romanian, Russian, Spanish, Turkish, Vietnamese

Table 3 Some broadly used open-source NMT toolkits

Toolkits	Language	Framework	URL
OpenNMT	Python	Pytorch	https://opennmt.net/
Fairseq	Python	Pytorch	https://github.com/facebookresearch/fairseq
Tensor2tensor	Python	Tensorflow	https://github.com/tensorflow/tensor2tensor
Sockeye	Python	Pytorch/MXNet	https://github.com/awslabs/sockeye
Marian	C++	*	https://marian-nmt.github.io/
THUMT	Python	PyTorch/TensorFlow	https://github.com/THUNLP-MT/THUMT
Hugging face	Python	PyTorch/TensorFlow	https://github.com/huggingface/transformers

former based pre-trained language models to generate more natural and precise answers.

6.2 Extensions to computer vision

Inspired by the success of Transformer in the field of NLP, researchers have applied Transformer to computer vision (CV) tasks to replace traditional CNNs[154–156]. Here, we briefly introduce the application of Transformer for image classification and image generation.

Transformer for image classification. Early studies incorporated the self-attention strategy on the basis of CNN-based architectures. Later, vision Transformer (ViT)[154], a pure architecture Transformer-based architecture, was proposed for image classification task. ViT first splits an image into a sequence of patches, which are treated as tokens in NLP. Then, patch embeddings are fed into a standard Transformer encoder and the hidden states are calculated. Similar to BERT, a special token "`class`" is added into the input sequence, and its corresponding hidden states are utilized to perform image classification. Following the paradigm of ViT, many ViT variants have been proposed, such as Transformer iN Transformer (TNT)[157], Swin Transformer[155], T2T-ViT[158], and CSWin Transformer[159].

Transformer for image generation. Transformers are also widely utilized in image generation task. Transgenerative adversarial networks (TransGAN)[160] presents a pure Transformer-based GAN architecture, where a memory-friendly Transformer-based generator with multiple stages could gradually increase the feature map resolution and a multi-scale Transformer-based discriminator could capture the semantic contexts and low-level textures simultaneously. Esser et al.[161] proposed a taming Transformer for image generation, where they use CNNs to learn a codebook of context-rich visual parts and utilize Transformers to model their composition within high-resolution images.

6.3 Extensions to audio tasks

Since Transformer has shown its great advantage in NLP and CV tasks, now it has also been extended to many audio-related tasks, such as audio classification[162], speech recognition[163, 164], and speech synthesis[165–167]. Here, we briefly introduce the Transformer for speech recognition and speech synthesis.

Transformer for speech recognition. Automatic speech recognition (ASR) can also be treated as a sequence-to-sequence task where the input is a speech sequence, and the output is its transcribed text sequence. Dong et al.[163] introduced the Transformer into speech recognition and proposed a speech-Transformer model. Tian et al.[164] proposed a synchronous transformer (Sync-Transformer) for online speech recognition. Sync-Transformer combines the Transformer and self-attention transducer. The self-attention in the Sync-Transformer encoder could make the node attend only its left context. Then its decoder starts to predict tokens immediately after a chunk of the state sequence is produced by the encoder.

Transformer for speech synthesis. Speech synthesis, or text-to-speech, aims at synthesizing natural speech given text and is a hot research topic. Li et al.[165] proposed a Transformer-based speech synthesis model, where the self-attention mechanism is introduced to replace the traditional RNNs in the encoder and decoder. Some subsequent studies are proposed to enhance the robustness of the Transformer-based speech synthesis model. Meanwhile, the Transformer-based pre-trained language model is also widely utilized in speech synthesis to enhance text representation[166, 167].

6.4 Extensions to multi-modal tasks

Due to its flexible and expansible architecture, Transformer has also been applied in many multi-modal scenarios. Here, we briefly introduce the application of Transformer for visual question answering and multi-modal pre-training.

Transformer for visual question answering. Visual question answering (VQA) aims at generating answers given a question in natural language and a corresponding image. Recently, VQA has attracted much attention and Transformers also play an important role in this task. Hu et al.[168] proposed a multi-modal Transformer for VQA, where they represent the semantic embeddings of different modalities into a common space and self-at-

tention is applied to model inter- and intra-modality context. Meanwhile, LaTr[169] is a layout-aware Transformer for scene-text VQA and consists of three building blocks: A Transformer-based language block which is pre-trained on only text, a layout-aware denoising pre-training block, and a ViT block to extract the visual features.

Transformer for multi-modal pre-training. Inspired by success of Transformer-based pre-trained language models (e.g., BERT), researchers have proposed various Transformer-based multi-modal pre-training models, such as SpeechBERT[170] to model audio and text, VisualBERT[171] and VL-BERT[172] to model vision and text, VideoBERT[173] to model the video and text. CLIP[174] is a contrastive language-image pre-training model, that takes natural language as a supervision to learn image representations. CLIP jointly trains an image encoder and a text encoder by utilizing a contrastive objective to predict the correct pairings of (image, text) in the training set. UniT[175] is a multi-modal multi-task pre-training model across different domains via a unified Transformer. UniT jointly learns 7 different visual and textual tasks including object detection, visual entailment, visual question answering, and natural language understanding tasks. Experimental results show that UniT could improve the performance of several multi-modal tasks.

In conclusion, Transformer has demonstrated its architectural superiority to model various NLP, CV and audio tasks. Furthermore, Transformer has shown its potential for building a general-purpose model to handle a vast number of multi-modal applications.

7 Current status and analysis

NMTs, especially the Transformer-based NMT, have proven their power for MT tasks. However, we should be soberly aware that NMT is still far from satisfactory. In this section, we briefly introduce the current status of NMT.

For formal texts in high-resource language pairs, the current NMT can produce high-quality translation results. It must be admitted that tremendous progress has been made by the Transformer-based NMT model. In the date-rich language pairs (such as Chinese-English and German-English), the translation quality of the formal written text (such as news, reports or daily conversations) by NMT models is quite high. Even the experimental results from Hassan et al.[8] in the WMT 2017 news Translation task from Chinese to English show that "the Translation quality is at human parity when compared to professional human translations". Due to its low cost, high efficiency, and high translation quality, NMT could basically meet the translation demands and has been widely used.

Low-resource translation is still far from satisfactory. Although many methods have been proposed for low-resource NMT, the quality of low-resource translation is still lower than that of high-resource translation. As a data-driven method, Transformer still heavily relies on parallel training data. For the low-resource translation, NMT is only helpful when the users just want to acquire the topic or general gist of the input sentences or documents.

Some language phenomena are still not well addressed. It is an important issue for an NMT model to deal with many special language phenomena, such as unseen words, entities and terminologies, abbreviations, idioms, omissions, and ambiguity. More seriously, language is dynamic and new language phenomena appear almost every day. However, the current NMT models still learn the translation knowledge from parallel training data and cannot handle these phenomena well. These problems occur widely and severely harm the translation quality of NMT.

The utilized information in multi-modal NMT is limited. Take the speech translation as an example. The current speech translation only incorporates speech signals. In face-to-face communication scenarios, in addition to the speech signals, some paralinguistic features, such as tone of voice, posture and gesture, are also important for the neural model to recognize the intention, desire and mood of speakers, while such paralinguistic features are not taken into consideration.

The improvement of pre-trained languages model for the NMT task is not that significant. Although many pre-training language models for NMT have been proposed to improve the translation quality, the improvement of the translation task is significantly less than that of other NLP tasks, such as text classifications, question answering, and named entity recognition. Potential evidence is that in the WMT 2021 Biomedical Translation Shared Task[176], in all eight participating teams, only two teams utilize the pre-trained language models. In contrast, the leaderboards of the general language understanding evaluation (GLUE) benchmark[11] and superGLUE[12] are almost various pre-training languages models.

Overall, Transformer-based NMT has brought the quality of NMT up to a new stage. For high-resource NMT with formal texts, the translation quality of the current NMT model can basically meet the translation demands for daily use. While in many scenarios, such as new language phenomena, multi-modal translation and low-resource translation, current Transformer-based NMT models still have a long way to go.

8 Future research directions

As we mentioned above, many NMT approaches have been proposed and have achieved remarkable progress,

[11] https://gluebenchmark.com/leaderboard

[12] https://super.gluebenchmark.com/leaderboard

while many challenges remain. In this section, we list some potential and important research directions.

Robust NMT. In addition to the formal written texts (such as news and reports), in practice, the NMT model faces various informal texts with different noises, such as speech noises from the ASR model, noisy web-crawled corpora or limited exposure to training data. Although many methods have been proposed to improve the robustness of NMT, there remains a huge gap between the translation quality with clean input and the real noise inputs. In addition, the current NMT models are also vulnerable to adversarial attacks, which are also a serious problem in commercial systems. Therefore, it remains a major challenge to design more robust NMT models to deal with informal texts with real noise inputs and adversarial attacks.

NMT with broader background knowledge. In many cases, only source sentences or speech cannot support the NMT model to produce the desired translation result. Meanwhile, some language phenomena, such as terminologies, abbreviations, idioms, and omissions, are still not well addressed. To overcome these problems, a potential solution is to combine broader background knowledge, which includes not only the mentioned syntactic structures, bilingual lexicons, phrase tables and terminologies, and knowledge graphs but also commonsense knowledge, local culture, history, etc. The knowledge is hard to learn from parallel sentence pairs. Accordingly, it remains a major challenge to represent and incorporate broader background knowledge into NMT.

Designing better pre-trained language models for low-resource NMT. Recently, the emergence of pre-trained language models has brought the NLP community to a new era. Although many methods are proposed to utilize the pre-trained language models for NMT, the improvement of the translation task is not as great as that of other natural language understanding tasks. Thus, designing better pre-trained language models or multilingual pre-trained language models for low-resource NMTs is an important future research direction.

Incorporating more information for multi-modal NMT scenarios. For speech translation, despite of large improvements, the end-to-end framework currently cannot perform on par with the cascaded method in many cases. Meanwhile, paralinguistic features, such as tone of voice, posture and gesture, also need to be taken into consideration. The current image-text translation only addresses the translation of image captions. In addition, scene image text translation is another important scenario, but it has attracted little attention. Scene image text refers to text in natural scenes, captured in its native environment, such as the scanned document, signboards and product packaging. Thus scene image text translation remains a major challenge due to complex backgrounds, various fonts, and imperfect imaging conditions.

Acknowledgements

This work was supported by Natural Science Foundation of China (Nos. 62006224 and 62122088).

Declarations of Conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, R. L. Mercer.- The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[2] P. Koehn, F. J. Och, D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of North American Chapter of Association for Computational Linguistics*, Edmonton, Canada, pp. 127–133, 2003.

[3] J. J. Zhang, C. Q. Zong. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, vol. 30, no. 5, pp. 16–25, 2015. DOI: 10.1109/MIS.2015.69.

[4] N. Kalchbrenner, P. Blunsom. Recurrent continuous translation models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Seattle, USA, pp. 1700–1709, 2013.

[5] I. Sutskever, O. Vinyals, Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 3104–3112, 2014.

[6] D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate. [Online], Available: https://arxiv.org/abs/1409.0473, 2015.

[7] Y. H. Wu, M. Schuster, Z. F. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. B. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean. Google′s neural machine translation system: Bridging the gap between human and machine translation. [Online], Available: https://arxiv.org/abs/1609.08144, 2016.

[8] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. D. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. J. Liu, T. Y. Liu, R. Q. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. J. Wu, S. Z. Wu, Y. C. Xia, D. D. Zhang, Z. R. Zhang, M. Zhou. Achieving human parity on automatic Chinese to English news translation. [Online], Available: https://arxiv.org/abs/1803.05567, 2018.

[9] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, pp. 1243–1252, 2017.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, Long Beach, USA, pp. 6000–6010, 2017.

[11] T. Luong, H. Pham, C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Lisbon, Portugal, pp. 1412–1421, 2015. DOI: 10.18653/v1/D15-1166.

[12] J. Gehring, M. Auli, D. Grangier, Y. Dauphin. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL, Vancouver, Canada, pp. 123–135, 2017. DOI: 10.18653/v1/P17-1012.

[13] R. Sennrich, B. Haddow, A. Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, Berlin, Germany, pp. 86–96, 2016. DOI: 10.18653/v1/P16-1009.

[14] A. Karakanta, J. Dehdari, J. van Genabith. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, vol. 32, no. 1, pp. 167–189, 2018. DOI: 10.1007/s10590-017-9203-5.

[15] S. Edunov, M. Ott, M. Auli, D. Grangier. Understanding back-translation at scale. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 489–500, 2018. DOI: 10.18653/v1/D18-1045.

[16] J. J. Zhang, C. Q. Zong. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Austin, USA, pp. 1535–1545, 2016. DOI: 10.18653/v1/D16-1160.

[17] Y. Cheng, W. Xu, Z. J. He, W. He, H. Wu, M. S. Sun, Y. Liu. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, Berlin, Germany, pp. 1965–1974, 2016. DOI: 10.18653/v1/P16-1185.

[18] D. He, Y. C. Xia, T. Qin, L. W. Wang, N. H. Yu, T. Y. Liu, W. Y. Ma. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain, pp. 820–828, 2016.

[19] M. Artetxe, G. Labaka, E. Agirre, K. Cho. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[20] G. Lample, A. Conneau, L. Denoyer, M. Ranzato. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[21] T. Mikolov, Q. V. Le, I. Sutskever. Exploiting similarities among languages for machine translation. [Online], Available: https://arxiv.org/abs/1309.4168, 2013.

[22] M. Zhang, Y. Liu, H. B. Luan, M. S. Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL, Vancouver, Canada, pp. 1959–1970, 2017. DOI: 10.18653/v1/P17-1179.

[23] M. Artetxe, G. Labaka, E. Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL, Melbourne, Australia, pp. 789–798, 2018. DOI: 10.18653/v1/P18-1073.

[24] T. Mohiuddin, S. Joty. Unsupervised word translation with adversarial autoencoder. *Computational Linguistics*, vol. 46, no. 2, pp. 257–288, 2020. DOI: 10.1162/coli_a_00374.

[25] M. Artetxe, G. Labaka, E. Agirre. Unsupervised statistical machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 3632–3642, 2018. DOI: 10.18653/v1/D18-1399.

[26] G. Lample, M. Ott, A. Conneau, L. Denoyer, M. Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 5039–5049, 2018. DOI: 10.18653/v1/D18-1549.

[27] S. Ren, Z. R. Zhang, S. J. Liu, M. Zhou, S. Ma. Unsupervised neural machine translation with SMT as posterior regularization. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference and 9th AAAI Symposium on Educational Advances in Artificial Intelligence*, Honolulu, USA, Article number 30, 2019. DOI: 10.1609/aaai.v33i01.3301241.

[28] M. Artetxe, G. Labaka, E. Agirre. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 194–203, 2019. DOI: 10.18653/v1/P19-1019.

[29] X. Garcia, A. Siddhant, O. Firat, A. Parikh. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, pp. 1126–1137, 2021. DOI: 10.18653/v1/2021.naacl-main.89.

[30] A. Üstün, A. Berard, L. Besacier, M. Gallé. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, pp. 6650–6662, 2021. DOI: 10.18653/v1/2021.emnlp-main.533.

[31] G. H. Chen, S. M. Ma, Y. Chen, L. Dong, D. D. Zhang, J. Pan, W. P. Wang, F. R. Wei. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, pp. 15–26, 2021. DOI: 10.18653/v1/2021.emnlp-main.2.

[32] G. H. Chen, S. M. Ma, Y. Chen, D. D. Zhang, J. Pan, W. P. Wang, F. R. Wei. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 142–157, 2022. DOI: 10.18653/v1/2022.acl-long.12.

[33] O. Firat, K. Cho, Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, San

Diego, USA, pp. 866–875, 2016. DOI: 10.18653/v1/N16-1101.

[34] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. H. Wu, Z. F. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. DOI: 10.1162/tacl_a_00065.

[35] R. Aharoni, M. Johnson, O. Firat. Massively multilingual neural machine translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Minneapolis, USA, pp. 3874–3884, 2019. DOI: 10.18653/v1/N19-1388.

[36] D. Sachan, G. Neubig. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, ACL, Brussels, Belgium, pp. 261–271, 2018. DOI: 10.18653/v1/W18-6327.

[37] G. Blackwood, M. Ballesteros, T. Ward. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, ACL, Santa Fe, USA, pp. 3112–3122, 2018.

[38] A. Bapna, O. Firat. Simple, scalable adaptation for neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, ACL, Hong Kong, China, pp. 1538–1548, 2019. DOI: 10.18653/v1/D19-1165.

[39] A. Eriguchi, S. F. Xie, T. Qin, H. Hassan. Building multilingual machine translation systems that serve arbitrary XY translations. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Seattle, USA, pp. 600–606, 2022. DOI: 10.18653/v1/2022.naacl-main.44.

[40] W. Y. Xie, Y. Feng, S. H. Gu, D. Yu. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL, pp. 5725–5737, 2021. DOI: 10.18653/v1/2021.acl-long.445.

[41] Z. H. Lin, L. W. Wu, M. X. Wang, L. Li. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL, pp. 293–305, 2021. DOI: 10.18653/v1/2021.acl-long.25.

[42] Q. Wang, J. J. Zhang. Parameter differentiation based multilingual neural machine translation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, the 34th AAAI Conference on Innovative Applications of Artificial Intelligence and the 12th IAAI Symposium on Educational Advances in Artificial Intelligence*, pp. 11440–11448, 2022.

[43] X. P. Qiu, T. X. Sun, Y. G. Xu, Y. F. Shao, N. Dai, X. J. Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020. DOI: 10.1007/s11431-020-1647-3.

[44] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, New Orleans, USA, pp. 2227–2237, 2018. DOI: 10.18653/v1/N18-1202.

[45] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Minneapolis, USA, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[46] S. Edunov, A. Baevski, M. Auli. Pre-trained language model representations for language generation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Minneapolis, USA, pp. 4052–4059, 2019. DOI: 10.18653/v1/N19-1409.

[47] J. H. Zhu, Y. C. Xia, L. J. Wu, D. He, T. Qin, W. G. Zhou, H. Q. Li, T. Y. Liu. Incorporating BERT into neural machine translation. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[48] J. C. Yang, M. X. Wang, H. Zhou, C. Q. Zhao, W. N. Zhang, Y. Yu, L. Li. Towards making the most of BERT in neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9378–9385, 2020. DOI: 10.1609/aaai.v34i05.6479.

[49] K. T. Song, X. Tan, T. Qin, J. F. Lu, T. Y. Liu. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, pp. 5926–5936, 2019.

[50] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Q. Zhou, W. Li, P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, vol. 21, no. 1, Article number 140, 2020.

[51] M. Lewis, Y. H. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 7871–7880, 2020. DOI: 10.18653/v1/2020.acl-main.703.

[52] W. X. Wang, W. X. Jiao, Y. C. Hao, X. Wang, S. M. Shi, Z. P. Tu, M. Lyu. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 2591–2600, 2022. DOI: 10.18653/v1/2022.acl-long.185.

[53] T. Pires, E. Schlinger, D. Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 4996–5001, 2019. DOI: 10.18653/v1/P19-1493.

[54] A. Conneau, G. Lample. Cross-lingual language model pretraining. In *Proceedings of the 33rd International*

Conference on Neural Information Processing Systems, Vancouver, Canada, Article number 634, 2019.

[55] Z. H. Lin, X. Pan, M. X. Wang, X. P. Qiu, J. T. Feng, H. Zhou, L. Li. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of Conference on Empirical Methods in Natural Language Processing* ACL, pp. 2649–2663, 2020. DOI: 10.18653/v1/2020.emnlp-main.210.

[56] Z. W. Chi, L. Dong, S. M. Ma, S. H. Huang, S. Singhal, X. L. Mao, H. Y. Huang, X. Song, F. R. Wei. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, pp. 1671–1683, 2021. DOI: 10.18653/v1/2021.emnlp-main.125.

[57] Y. H. Liu, J. T. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020. DOI: 10.1162/tacl_a_00343.

[58] P. F. Li, L. Y. Li, M. Zhang, M. H. Wu, Q. Liu. Universal conditional masked language pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 6379–6391, 2022. DOI: 10.18653/v1/2022.acl-long.442.

[59] R. Bawden, R. Sennrich, A. Birch, B. Haddow. Evaluating discourse phenomena in neural machine translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, New Orleans, USA, pp. 1304–1313, 2018. DOI: 10.18653/v1/N18-1118.

[60] E. Voita, P. Serdyukov, R. Sennrich, I. Titov. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL, Melbourne, Australia, pp. 1264–1274, 2018. DOI: 10.18653/v1/P18-1117.

[61] J. C. Zhang, H. B. Luan, M. S. Sun, F. F. Zhai, J. F. Xu, M. Zhang, Y. Liu. Improving the transformer translation model with document-level context. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 533–542, 2018. DOI: 10.18653/v1/D18-1049.

[62] B. Zhang, A. Bapna, M. Johnson, A. Dabirmoghaddam, N. Arivazhagan, O. Firat. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 4176–4192, 2022. DOI: 10.18653/v1/2022.acl-long.287.

[63] Z. P. Tu, Y. Liu, S. M. Shi, T. Zhang. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 407–420, 2018. DOI: 10.1162/tacl_a_00029.

[64] E. Voita, R. Sennrich, I. Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 1198–1212, 2019. DOI: 10.18653/v1/P19-1116.

[65] E. Voita, R. Sennrich, I. Titov. Context-aware monolingual repair for neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, ACL, Hong Kong, China, pp. 877–886, 2019. DOI: 10.18653/v1/D19-1081.

[66] L. Lupo, M. Dinarelli, L. Besacier. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 4557–4572, 2022. DOI: 10.18653/v1/2022.acl-long.312.

[67] S. H. Kuang, D. Y. Xiong, W. H. Luo, G. D. Zhou. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, ACL, Santa Fe, USA, pp. 596–606, 2018.

[68] X. M. Kang, Y. Zhao, J. J. Zhang, C. Q. Zong. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, pp. 2242–2254, 2020. DOI: 10.18653/v1/2020.emnlp-main.175.

[69] P. Y. Huang, F. Liu, S. R. Shiang, J. Oh, C. Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers*, ACL, Berlin, Germany, pp. 639–645, 2016. DOI: 10.18653/v1/W16-2360.

[70] I. Calixto, M. Rios, W. Aziz. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 6392–6405, 2019. DOI: 10.18653/v1/P19-1642.

[71] J. Ive, P. Madhyastha, L. Specia. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 6525–6538, 2019. DOI: 10.18653/v1/P19-1653.

[72] X. Huang, J. J. Zhang, C. Q. Zong. Entity-level cross-modal learning improves multi-modal machine translation. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP*, ACL, Punta Cana, Dominican Republic, pp. 1067–1080, 2021. DOI: 10.18653/v1/2021.findings-emnlp.92.

[73] D. X. Wang, D. Y. Xiong. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, Palo Alto, USA, pp. 2720–2728, 2021. DOI: 10.1609/aaai.v35i4.16376.

[74] B. Li, C. H. Lv, Z. F. Zhou, T. Zhou, T. Xiao, A. X. Ma, J. B. Zhu. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 6327–6337, 2022. DOI: 10.18653/v1/2022.acl-long.438.

[75] A. Bérard, O. Pietquin, L. Besacier, C. Servan. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *Proceedings of the NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, Barcelona, Spain, 2016. [Online], Available: https://hal.science/hal-01408086.

[76] R. J. Weiss, J. Chorowski, N. Jaitly, Y. H. Wu, Z. F.

Chen. Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of Interspeech*, Stockholm, Sweden, pp. 2625–2629, 2017. DOI: 10.21437/Interspeech.2017-503.

[77] R. Ye, M. X. Wang, L. Li. Cross-modal contrastive learning for speech translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Seattle, USA, pp. 5099–5113, 2022. DOI: 10.18653/v1/2022.naacl-main.376.

[78] G. Sant, G. I. Gállego, B. Alastruey, M. R. Costa-Jussà. Multiformer: A head-configurable transformer-based model for direct speech translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, ACL, Seattle, USA, pp. 277–284, 2022. DOI: 10.18653/v1/2022.naacl-srw.34.

[79] T. K. Lam, S. Schamoni, S. Riezler. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 245–254, 2022. DOI: 10.18653/v1/2022.acl-short.27.

[80] S. Bansal, H. Kamper, K. Livescu, A. Lopez, S. Goldwater. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Minneapolis, USA, pp. 58–68, 2019. DOI: 10.18653/v1/N19-1006.

[81] C. Y. Wang, Y. Wu, S. J. Liu, Z. L. Yang, M. Zhou. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9161–9168, 2020. DOI: 10.1609/aaai.v34i05.6452.

[82] S. Y. Chen, C. Y. Wang, Z. Y. Chen, Y. Wu, S. J. Liu, Z. Chen, J. Y. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. M. Qian, Y. Qian, J. Wu, M. Zeng, X. Z. Yu, F. R. Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022. DOI: 10.1109/JSTSP.2022.3188113.

[83] Y. Tang, H. Y. Gong, N. Dong, C. H. Wang, W. N. Hsu, J. T. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli, J. Pino. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 1488–1499, 2022. DOI: 10.18653/v1/2022.acl-long.105.

[84] Y. C. Liu, H. Xiong, J. J. Zhang, Z. J. He, H. Wu, H. F. Wang, C. Q. Zong. End-to-end speech translation with knowledge distillation. *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 1128–1132, 2019.

[85] H. Inaguma, T. Kawahara, S. Watanabe. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, pp. 1872–1881, 2021. DOI: 10.18653/v1/2021.naacl-main.150.

[86] Y. Tang, J. Pino, X. Li, C. H. Wang, D. Genzel. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL, pp. 4252–4261, 2021. DOI: 10.18653/v1/2021.acl-long.328.

[87] Y. Ren, J. L. Liu, X. Tan, C. Zhang, T. Qin, Z. Zhao, T. Y. Liu. SimulSpeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 3787–3796, 2020. DOI: 10.18653/v1/2020.acl-main.350.

[88] E. Salesky, M. Sperber, A. W. Black. Exploring phoneme-level speech representations for end-to-end speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 1835–1841, 2019. DOI: 10.18653/v1/P19-1179.

[89] E. Salesky, A. W. Black. Phone features improve speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 2388–2397, 2020. DOI: 10.18653/v1/2020.acl-main.217.

[90] J. T. Gu, J. Bradbury, C. M. Xiong, V. O. K. Li, R. Socher. Non-autoregressive neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018. [Online], Available: https://openreview.net/pdf?id=B1l8BtlCb.

[91] X. W. Zhang, J. S. Su, Y. Qin, Y. Liu, R. R. Ji, H. J. Wang. Asynchronous bidirectional decoding for neural machine translation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, USA, Article number 699, 2018.

[92] L. Zhou, J. J. Zhang, C. Q. Zong. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 91–105, 2019. DOI: 10.1162/tacl_a_00256.

[93] J. J. Zhang, L. Zhou, Y. Zhao, C. Q. Zong. Synchronous bidirectional inference for neural sequence generation. *Artificial Intelligence*, vol. 281, Article number 103234, 2020. DOI: 10.1016/j.artint.2020.103234.

[94] Y. R. Wang, F. Tian, D. He, T. Qin, C. X. Zhai, T. Y. Liu. Non-autoregressive machine translation with auxiliary regularization. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence*, Honolulu, USA, Article number 659, 2019. DOI: 10.1609/aaai.v33i01.33015377.

[95] L. Zhou, J. J. Zhang, Y. Zhao, C. Q. Zong. Non-autoregressive neural machine translation with distortion model. In *Proceedings of the 9th CCF International Conference on Natural Language Processing and Chinese Computing*, Springer, Zhengzhou, China, pp. 403–415, 2020. DOI: 10.1007/978-3-030-60450-9_32.

[96] L. Ding, L. Y. Wang, S. M. Shi, D. C. Tao, Z. P. Tu. Redistributing low-frequency words: Making the most of monolingual data in non-autoregressive translation. In *Proceedings of the 60th Annual Meeting of the Associ-*

ation for Computational Linguistics, ACL, Dublin, Ireland, pp. 2417–2426, 2022. DOI: 10.18653/v1/2022.acl-long.172.

[97] C. Z. Shao, X. F. Wu, Y. Feng. One reference is not enough: Diverse distillation with reference selection for non-autoregressive translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Seattle, USA, pp. 3779–3791, 2022. DOI: 10.18653/v1/2022.naacl-main.277.

[98] C. Q. Wang, J. Zhang, H. Q. Chen. Semi-autoregressive neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 479–488, 2018. DOI: 10.18653/v1/D18-1044.

[99] M. Ghazvininejad, O. Levy, Y. H. Liu, L. Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, ACL, Hong Kong, China, pp. 6112–6121, 2019. DOI: 10.18653/v1/D19-1633.

[100] J. T. Gu, C. H. Wang, J. K. Zhao. Levenshtein transformer. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019.

[101] M. H. Zhu, J. L. Wang, C. G. Yan. Non-autoregressive neural machine translation with consistency regularization optimized variational framework. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Seattle, USA, pp. 607–617, 2022. DOI: 10.18653/v1/2022.naacl-main.45.

[102] J. Lee, E. Mansimov, K. Cho. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 1173–1182, 2018. DOI: 10.18653/v1/D18-1149.

[103] C. Zeng, J. J. Chen, T. Y. Zhuang, R. Xu, H. Yang, Q. Ying, S. M. Tao, Y. H. Xiao. Neighbors are not strangers: Improving non-autoregressive translation under low-frequency lexical constraints. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Seattle, USA, pp. 5777–5790, 2022. DOI: 10.18653/v1/2022.naacl-main.424.

[104] R. Sennrich, B. Haddow, A. Birch. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers*, ACL, Berlin, Germany, pp. 371–376, 2016. DOI: 10.18653/v1/W16-2323.

[105] Y. C. Liu, L. Zhou, Y. N. Wang, Y. Zhao, J. J. Zhang, C. Q. Zong. A comparable study on model averaging, ensembling and reranking in NMT. In *Proceedings of the 7th CCF International Conference on Natural Language Processing and Chinese Computing*, Springer, Hohhot, China, pp. 299–308, 2018. DOI: 10.1007/978-3-319-99501-4_26.

[106] L. M. Liu, M. Utiyama, A. Finch, E. Sumita. Agreement on target-bidirectional neural machine translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, San Diego, USA, pp. 411–416, 2016. DOI: 10.18653/v1/N16-1046.

[107] Z. R. Zhang, S. Z. Wu, S. J. Liu, M. Li, M. Zhou, T. Xu. Regularizing neural machine translation by target-bidirectional agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 443–450, 2019. DOI: 10.1609/aaai.v33i01.3301443.

[108] J. S. Su, X. W. Zhang, Q. Lin, Y. Qin, J. F. Yao, Y. Liu. Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding. *Artificial Intelligence*, vol. 277, Article number 103168, 2019. DOI: 10.1016/j.artint.2019.103168.

[109] L. Zhou, J. J. Zhang, C. Q. Zong, H. Yu. Sequence generation: From both sides to the middle. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 5471–5477, 2019.

[110] P. Arthur, G. Neubig, S. Nakamura. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Austin, USA, pp. 1557–1567, 2016. DOI: 10.18653/v1/D16-1162.

[111] Y. Feng, S. Y. Zhang, A. D. Zhang, D. Wang, A. Abel. Memory-augmented neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Copenhagen, Denmark, pp. 1390–1399, 2017. DOI: 10.18653/v1/D17-1146.

[112] J. C. Zhang, Y. Liu, H. B. Luan, J. F. Xu, M. S. Sun. Prior knowledge integration for neural machine translation using posterior regularization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL, Vancouver, Canada, pp. 1514–1523, 2017. DOI: 10.18653/v1/P17-1139.

[113] Y. Zhao, J. J. Zhang, Z. J. He, C. Q. Zong, H. Wu. Addressing troublesome words in neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Brussels, Belgium, pp. 391–400, 2018. DOI: 10.18653/v1/D18-1036.

[114] M. T. Luong, C. D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, Berlin, Germany, pp. 1054–1063, 2016. DOI: 10.18653/v1/P16-1100.

[115] R. Sennrich, B. Haddow, A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, Berlin, Germany, pp. 1715–1725, 2016. DOI: 10.18653/v1/P16-1162.

[116] X. Wang, Z. P. Tu, D. Y. Xiong, M. Zhang. Translating phrases in neural machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Copenhagen, Denmark, pp. 1421–1431, 2017. DOI: 10.18653/v1/D17-1149.

[117] L. Dahlmann, E. Matusov, P. Petrushkov, S. Khadivi. Neural machine translation leveraging phrase-based models in a hybrid search. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, ACL, Copenhagen, Denmark, pp. 1411–1420, 2017. DOI: 10.18653/v1/D17-1148.

[118] Y. Zhao, Y. N. Wang, J. J. Zhang, C. Q. Zong. Phrase table as recommendation memory for neural machine translation. In *Proceedings of International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 4609–

4615, 2018.

[119] H. F. Xu, J. van Genabith, D. Y. Xiong, Q. H. Liu, J. Y. Zhang. Learning source phrase representations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 386–396, 2020. DOI: 10.18653/v1/2020.acl-main.37.

[120] M. Huck, V. Hangya, A. Fraser. Better OOV translation with bilingual terminology mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 5809–5815, 2019. DOI: 10.18653/v1/P19-1581.

[121] G. Dinu, P. Mathur, M. Federico, Y. Al-Onaizan. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 3063–3068, 2019. DOI: 10.18653/v1/P19-1294.

[122] Y. Zhao, J. J. Zhang, Y. Zhou, C. Q. Zong. Knowledge graphs enhanced neural machine translation. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pp. 4039–4045, 2021. DOI: 10.24963/ijcai.2020/559.

[123] Y. Zhao, L. Xiang, J. N. Zhu, J. J. Zhang, Y. Zhou, C. Q. Zong. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, ACL, Barcelona, Spain, pp. 4495–4505, 2020. DOI: 10.18653/v1/2020.coling-main.397.

[124] J. J. Hu, H. Hayashi, K. Cho, G. Neubig. DEEP: DEnoising entity pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 1753–1766, 2022. DOI: 10.18653/v1/2022.acl-long.123.

[125] J. Tiedemann, S. Thottingal. OPUS-MT-Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal, pp. 479–480, 2020.

[126] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand, pp. 79–86, 2005.

[127] D. Elliott, S. Frank, K. Sima'an, L. Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, ACL, Berlin, Germany, pp. 70–74, 2016. DOI: 10.18653/v1/W16-3210.

[128] A. C. Kocabiyikoglu, L. Besacier, O. Kraif. Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 2018.

[129] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, M. Turchi. MuST-C: A multilingual speech translation corpus. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Minneapolis, USA, pp. 2012–2017, 2019. DOI: 10.18653/v1/N19-1202.

[130] M. Cettolo, C. Girardi, M. Federico. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, Trento, Italy, pp. 261–268, 2012.

[131] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, S. Khudanpur. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation*, Heidelberg, Germany, 2013. [Online], Available: https://aclanthology.org/2013.iwslt-papers.14/.

[132] J. J. Zhang, C. Q. Zong. Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, vol. 63, no. 10, pp. 2028–2050, 2020. DOI: 10.1007/s11431-020-1632-x.

[133] F. Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020. DOI: 10.1613/jair.1.12007.

[134] Z. X. Tan, S. Wang, Z. H. Yang, G. Chen, X. C. Huang, M. S. Sun, Y. Liu. Neural machine translation: A review of methods, resources, and tools. *AI Open*, vol. 1, pp. 5–21, 2020. DOI: 10.1016/j.aiopen.2020.11.001.

[135] T. X. Sun, X. Y. Liu, X. P. Qiu, X. J. Huang. Paradigm shift in natural language processing. *Machine Intelligence Research*, vol. 19, no. 3, pp. 169–183, 2022. DOI: 10.1007/s11633-022-1331-6.

[136] Y. H. Liu, M. Ott, N. Goyal, J. F. Du, M. Joshi, D. Q. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. [Online], Available: https://arxiv.org/abs/1907.11692, 2019.

[137] M. Joshi, D. Q. Chen, Y. H. Liu, D. S. Weld, L. Zettlemoyer, O. Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020. DOI: 10.1162/tacl_a_00300.

[138] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever. Improving language understanding by generative pre-training. [Online], Available: https://openai.com/research/language-unsupervised, Nov. 7, 2022.

[139] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, vol. 1, no. 8, Article number 9, 2019.

[140] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 159, 2020.

[141] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, N. Shazeer. Generating Wikipedia by summarizing long sequences. In *Proceedings of the 6th International Conference on Learning Representations*,Vancouver, Canada, 2018.

[142] W. Xiao, I. Beltagy, G. Carenini, A. Cohan. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Lin-*

guistics, ACL, Dublin, Ireland, pp. 5245–5263, 2022. DOI: 10.18653/v1/2022.acl-long.360.

[143] Y. Rui, V. I. S. Carmona, M. Pourvali, Y. Xing, W. W. Yi, H. B. Ruan, Y. Zhang. Knowledge mining: A cross-disciplinary survey. *Machine Intelligence Research*, vol. 19, no. 2, pp. 89–114, 2022. DOI: 10.1007/s11633-022-1323-6.

[144] A. Saxena, A. Kochsiek, R. Gemulla. Sequence-to-sequence knowledge graph completion and question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 2814–2828, 2022. DOI: 10.18653/v1/2022.acl-long.201.

[145] T. X. Sun, Y. F. Shao, X. P. Qiu, Q. P. Guo, Y. R. Hu, X. J. Huang, Z. Zhang. CoLAKE: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, pp. 3660–3670, 2020. DOI: 10.18653/v1/2020.coling-main.327.

[146] M. Henderson, I. Vulić, D. Gerz, I. Casanueva, P. Budzianowski, S. Coope, G. Spithourakis, T. H. Wen, N. Mrkšić, P. H. Su. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 5392–5404, 2019. DOI: 10.18653/v1/P19-1536.

[147] Y. Z. Zhang, S. Q. Sun, M. Galley, Y. C. Chen, C. Brockett, X. Gao, J. F. Gao, J. J. Liu, B. Dolan. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL, pp. 270–278, 2020. DOI: 10.18653/v1/2020.acl-demos.30.

[148] Z. Y. Ma, J. J. Li, G. H. Li, Y. J. Cheng. UniTranSeR: A unified transformer semantic representation framework for multimodal task-oriented dialog system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL, Dublin, Ireland, pp. 103–114, 2022. DOI: 10.18653/v1/2022.acl-long.9.

[149] H. B. Bao, L. Dong, F. R. Wei, W. H. Wang, N. Yang, L. Cui, S. H. Piao, M. Zhou. Inspecting unification of encoding and matching with transformer: A case study of machine reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, ACL, Hong Kong, China, pp. 14–18, 2019. DOI: 10.18653/v1/D19-5802.

[150] Z. W. Bai, J. P. Liu, M. Q. Wang, C. X. Yuan, X. J. Wang. Exploiting diverse information in pre-trained language model for multi-choice machine reading comprehension. *Applied Sciences*, vol. 12, no. 6, Article number 3072, 2022. DOI: 10.3390/app12063072.

[151] K. Nishida, I. Saito, K. Nishida, K. Shinoda, A. Otsuka, H. Asano, J. Tomita. Multi-style generative reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, Florence, Italy, pp. 2273–2284, 2019. DOI: 10.18653/v1/P19-1220.

[152] C. Zhao, C. Y. Xiong, C. Rosset, X. Song, P. N. Bennett, S. Tiwary. Transformer-XH: Multi-evidence reasoning with extra hop attention. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[153] G. Izacard, E. Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, pp. 874–880, 2021. DOI: 10.18653/v1/2021.eacl-main.74.

[154] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[155] Z. Liu, Y. T. Lin, Y. Cao, H. Hu, Y. X. Wei, Z. Zhang, S. Lin, B. N. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 10012–10022, 2021. DOI: 10.1109/ICCV48922.2021.00986.

[156] X. Q. Zhang, R. H. Jiang, C. X. Fan, T. Y. Tong, T. Wang, P. C. Huang. Advances in deep learning methods for visual tracking: Literature review and fundamentals. *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 311–333, 2021. DOI: 10.1007/s11633-020-1274-8.

[157] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang. Transformer in transformer. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, pp. 15908–15919, 2021.

[158] L. Yuan, Y. P. Chen, T. Wang, W. H. Yu, Y. J. Shi, Z. H. Jiang, F. E. H. Tay, J. S. Feng, S. C. Yan. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 558–567, 2021. DOI: 10.1109/ICCV48922.2021.00060.

[159] X. Y. Dong, J. M. Bao, D. D. Chen, W. M. Zhang, N. H. Yu, L. Yuan, D. Chen, B. N. Guo. CSWin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 12124–12134, 2022. DOI: 10.1109/CVPR52688.2022.01181.

[160] Y. F. Jiang, S. Y. Chang, Z. Y. Wang. TransGAN: Two pure transformers can make one strong GAN, and that can scale up. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, pp. 14745–14758, 2021.

[161] P. Esser, R. Rombach, B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 12873–12883, 2021. DOI: 10.1109/CVPR46437.2021.01268.

[162] Y. Gong, Y. A. Chung, J. Glass. AST: Audio spectrogram transformer. [Online], Available: https://arxiv.org/abs/2104. 01778, 2021.

[163] L. H. Dong, S. Xu, B. Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, pp. 5884–5888, 2018. DOI: 10.1109/ICASSP.2018.8462506.

[164] Z. K. Tian, J. Y. Yi, Y. Bai, J. H. Tao, S. Zhang, Z. Q.

Wen. Synchronous transformers for end-to-end speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 7884–7888, 2020. DOI: 10.1109/ICASSP40776.2020.9054260.

[165] N. H. Li, S. J. Liu, Y. Q. Liu, S. Zhao, M. Liu. Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 6706–6713, 2019. DOI: 10.1609/aaai.v33i01.33016706.

[166] Y. Jia, H. G. Zen, J. Shen, Y. Zhang, Y. H. Wu. PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, pp. 151–155, 2021.

[167] G. H. Xu, W. Song, Z. C. Zhang, C. Zhang, X. D. He, B. W. Zhou. Improving prosody modelling with cross-utterance Bert embeddings for end-to-end speech synthesis. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, pp. 6079–6083, 2021. DOI: 10.1109/ICASSP39728.2021.9414102.

[168] R. H. Hu, A. Singh, T. Darrell, M. Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textVQA. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9989–9999, 2020. DOI: 10.1109/CVPR42600.2020.01001.

[169] A. F. Biten, R. Litman, Y. S. Xie, S. Appalaraju, R. Manmatha. LaTr: Layout-aware transformer for scene-text VQA. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 16527–16537, 2022. DOI: 10.1109/CVPR52688.2022.01605.

[170] Y. S. Chuang, C. L. Liu, H. Y. Lee, L. S. Lee. SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, pp. 4168–4172, 2020.

[171] L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh, K. W. Chang. VisualBERT: A simple and performant baseline for vision and language. [Online], Available: https://arxiv.org/abs/1908.03557, 2019.

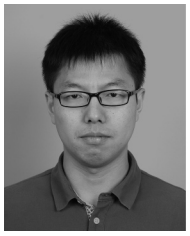
[172] W. J. Su, X. Z. Zhu, Y. Cao, B. Li, L. W. Lu, F. R. Wei, J. F. Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[173] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid. VideoBERT: A joint model for video and language representation learning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 7463–7472, 2019. DOI: 10.1109/ICCV.2019.00756.

[174] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

[175] R. H. Hu, A. Singh. UniT: Multimodal multitask learning with a unified transformer. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 1419–1429, 2021. DOI: 10.1109/ICCV48922.2021.00147.

[176] L. Yeganova, D. Wiemann, M. Neves, F. Vezzani, A. Siu, I. J. Unanue, M. Oronoz, N. Mah, A. Névéol, D. Martinez, R. Bawden, G. M. Di Nunzio, R. Roller, P. Thomas, C. Grozea, O. Perez-de-Viñaspre, M. V. Navarro, A. J. Yepes. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the 6th Conference on Machine Translation*, ACL, pp. 664–683, 2021.



Yang Zhao received the Ph. D. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences, China in 2019. He is currently an associate professor with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China.

His research interests include machine translation and natural language processing.

E-mail: yang.zhao@nlpr.ia.ac.cn (Corresponding author)
ORCID iD: 0000-0003-1028-3406

Jiajun Zhang received the Ph. D. degree in computer science from Institute of Automation, Chinese Academy of Sciences, China in 2011. He is currently a professor with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China.

His research interests include machine translation, multilingual natural language processing and deep learning.

E-mail: jjzhang@nlpr.ia.ac.cn

Chengqing Zong received the Ph. D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, China in 1998. He is currently a professor with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. He is a member of the International Committee on Computational Linguistics and the President of Asian Federation of Natural Language Processing. He is an Associate Editor for the *ACM Transactions on Asian and Low-resource Language Information Processing* and an Editorial Board Member of the *IEEE Intelligent Systems*, the journal *Machine Translation*, and the *Journal of Computer Science and Technology*. He served ACL-IJCNLP 2015 as the PC Co-Chair, IJCAI 2017, IJCAI-ECAI 2018, and AAAI 2019 as the Area Chair, and IJCNLP 2017 as the General Chair.

His research interests include natural language processing, machine translation and sentiment analysis.

E-mail: cqzong@nlpr.ia.ac.cn