Engineering 18 (2022) 143–153

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

# Progress in Machine Translation

Haifeng Wang [a,*], Hua Wu [a], Zhongjun He [a], Liang Huang [b], Kenneth Ward Church [b]

[a] Baidu Inc., Beijing 100193, China
[b] Baidu Research, Sunnyvale, CA 94089, USA

ARTICLE INFO

*Article history:*
Received 15 November 2020
Revised 30 January 2021
Accepted 29 March 2021
Available online 14 July 2021

*Keywords:*
Machine translation
Neural machine translation
Simultaneous translation

ABSTRACT

After more than 70 years of evolution, great achievements have been made in machine translation. Especially in recent years, translation quality has been greatly improved with the emergence of neural machine translation (NMT). In this article, we first review the history of machine translation from rule-based machine translation to example-based machine translation and statistical machine translation. We then introduce NMT in more detail, including the basic framework and the current dominant framework, Transformer, as well as multilingual translation models to deal with the data sparseness problem. In addition, we introduce cutting-edge simultaneous translation methods that achieve a balance between translation quality and latency. We then describe various products and applications of machine translation. At the end of this article, we briefly discuss challenges and future research directions in this field.



## 1. A brief history of machine translation (MT)

MT is the study of how to use computers to translate from one language into another. The concept of MT was first put forward by Warren Weaver in 1947 [1], just one year after the first computer, electronic numerical integrator and computer, was developed. From then on, MT has been considered to be one of the most challenging tasks in the field of natural language processing (NLP).

In terms of methodology, approaches to MT mainly fall into two categories: rule-based methods and corpus-based methods. From the time when the idea of MT was first proposed until the 1990s, rule-based methods were dominant. Rule-based machine translation (RBMT) methods use bilingual dictionaries and manually written rules to translate source language texts into target language texts. However, manually writing rules is labor intensive. Furthermore, rules are difficult to maintain and difficult to transfer from one domain to another, and from one language to another. Thus, it is difficult for rule-based systems to be scalable for open-domain translation and multilingual translation. At the very beginning, MT systems were mainly designed for military applications. In 1954, Georgetown University, with the cooperation of the now well-known computer manufacturer International Business Machines (IBM) Corporation, completed a Russian–English MT experiment for the first time using the IBM-701 computer, demonstrating that the dream of MT had become true. MT was a hot topic for more than a decade after the 1954 demonstration, but the boom ended abruptly with the Automatic Language Processing Advisory Committee (ALPAC) report in 1966 [2]. After the report, which was very skeptical of MT and led to a drastic cut in funding for MT research, it became extremely difficult to work on MT. The dominant scientific society today, the Association for Computational Linguistics (ACL), was originally named the Association for Machine Translation and Computational Linguistics in 1962, during the boom; however, it dropped the "MT" from its name in 1968, during the bust after the ALPAC report. Meanwhile, MT researchers continued to attempt to improve translation quality. In 1965, NLP researchers held the first International Conference on Computational Linguistics, which focused on rule-based parsing and translation. Starting in the 1970s, RBMT methods became more mature. In 1978, SYSTRAN, one of the first MT companies, launched a commercial translation system, which was one of the best-known examples of a commercially successful rule-based system at that time. Google used the service of SYSTRAN until 2007.

With the availability of bilingual corpora, corpus-based methods became dominant after the 2000s. There are three corpus-based MT methods: example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT). In the mid-1980s, EBMT was proposed to translate

* Corresponding author.
*E-mail address:* wanghaifeng@baidu.com (H. Wang).

https://doi.org/10.1016/j.eng.2021.03.023
2095-8099/

source texts by retrieving similar sentence pairs from the bilingual corpus [3]. The translation results from EBMT methods are of high quality if similar sentence pairs can be retrieved. However, EBMT methods have low coverage of translations because the bilingual corpora cannot cover all the linguistic phenomena of the language pairs. As a result, EBMT methods are usually used in computer-aided translation systems.

In 1990, Brown et al. [4] proposed the idea of SMT, in which machines automatically learn translation knowledge from a large amount of data instead of relying on human experts to write rules. The idea was more formally formulated as five different SMT models in 1993 [5]. SMT methods were not widely adopted at that time due to their complexity and the dominance of RBMT in commercial applications during the 1980s and 1990s. However, with the emergence of statistical methods, another NLP Conference—the Empirical Methods in Natural Language Processing Conference—began in 1996, with the aim of bringing together empirical methods from a range of different disciplines, including corpus-based methods from linguistics and information theory from engineering [6]. In 1999, researchers held a summer workshop at Johns Hopkins University [7], at which they reproduced five IBM models and released an SMT toolkit named Egypt, which greatly reduced the threshold of SMT. The word-based SMT toolkits GIZA and GIZA++ were subsequently released [8]. In 2003, phrase-based SMT methods were proposed [9], which further improved the translation quality. Based on phrase-based SMT methods, open-source systems such as "Pharaoh" and its upgraded version, "Moses," were released [10], greatly promoting the development of SMT systems. After that, SMT methods were widely adopted because of these available toolkits. In 2006, Google launched its internet translation service based on phrase-based SMT methods. Other companies such as Microsoft and Baidu also launched translation services in the years that followed. It should be noted that it is difficult for a single model to deal with various translation requests; thus, practical systems usually use hybrid methods [11] that integrate different MT models in order to improve translation performance. Encouraged by the success of SMT models, many researchers proposed novel models to further improve the performance of SMT methods, including factored SMT models [12] in which morphological information was introduced, hierarchical SMT models [13], and syntax-based SMT models with parsing trees on the source side and/or target side [14–17].

Although the use of SMT methods greatly improved the translation quality, such methods employ log-linear models to integrate multiple manually designed components such as a translation model, a language model, and a reordering model, which usually results in a serious reordering problem for distant language pairs. With strong progress in deep learning technology in speech, vision, and other fields, researchers began to apply deep learning technology to MT. In 2014, Bahdanau et al. [18] and Sutskever et al. [19] proposed end-to-end neural network translation models and formally used the term "neural machine translation." The basic idea of NMT is to map the source language into a dense semantic representation, and then generate the translation by using an attention mechanism. At the same time, Dong et al. [20] proposed a multilingual translation framework based on NMT, which is considered to be a breakthrough paper for multilingual translation in the history of NMT. In 2015, Baidu deployed the first large-scale NMT system in the world [21]. In 2016, Google also launched an NMT system [22], which was followed by other companies releasing their NMT systems. Thus, it only took about one year for NMT to be deployed online since it was first proposed in 2014, while it took about 16 years for SMT systems to be applied to online service. After that, a convolutional sequence-to-sequence translation model [23] and the Transformer model [24] were proposed, which again significantly improved the translation quality. This great improvement triggered a wide-ranging discussion on whether MT is as good as human translation. The great success of NMT has attracted many researchers who have developed various methods such as non-autoregressive models [25,26], unsupervised NMT models [27,28], and pretraining models on NMT [29], with the aim of improving multilingual translation quality and translation efficiency.

The great improvements that have been achieved in both speech technologies and MT have led to simultaneous translation (ST) as another promising direction for MT. Exploration in spoken language translation or speech translation began with a small experimental automatic interpreting system that was demonstrated at the International Telecommunication Union expo in 1983 [30]. Subsequently, a speech-to-speech (S2S) translation system called SpeechTrans was developed in 1988 [31], and was considered to be an important landmark system in speech translation [32]. In the following two decades, particularly since the establishment of the Consortium for Speech Translation Advanced Research in 1991, impressive speech translation systems have been developed, from domain-limited and vocabulary-limited systems [33–35] to open-domain spontaneous translations [36–40]. Meanwhile, the International Workshop on Spoken Language Translation (IWSLT) was organized in 2004, which again promoted the development of speech translation systems [39].

With the emergence of NMT and neural speech recognition, new ST systems are intended to automate simultaneous interpreting, in which the translation system interprets concurrently with the source-language speech, with a delay of only a few seconds. Simultaneous interpretation is extremely challenging and exhausting for humans, as it requires extreme concentration and skill to listen to and comprehend one language while speaking another. Thus, there are a limited number of qualified simultaneous interpreters worldwide. Furthermore, simultaneous interpreters usually work in teams of two or more and swap places every 15–30 min to prevent the error rate from growing exponentially [38]. Moreover, limited memory forces human interpreters to routinely omit source content [41]. Therefore, there is a critical need to develop simultaneous MT techniques to reduce the burden of human interpreters and to make simultaneous interpreting services more accessible and affordable. To this end, as an early work, Wang et al. [42] proposed a neural network-based method to split streaming speech into appropriate segments in order to improve speech translation quality. Ma et al. [43] developed an extremely simple but effective "prefix-to-prefix" framework that is tailored to the simultaneity requirement. This technique achieved controllable latency for the first time and rejuvenated the NLP community's interest in ST. Since then, many major research laboratories (Google, Microsoft, Facebook, Huawei, etc.) have joined the research in this direction, and commercial products from companies such as Baidu have been serving hundreds of conferences. This renewed interest resulted in the First Workshop on Automatic Simultaneous Translation being held at ACL 2020 and a new ST track at the International Conference on Spoken Language Translation (IWSLT) 2020.

## 2. Neural machine translation

There has been great improvement in NMT in recent years [44,45]. A typical NMT model contains two components: An encoder network maps the source sentence into a real-valued vector, from which a decoder network produces the translation. This process is analogous to a human's translation. The NMT model first "reads" the whole source sentence; then, based on its understanding of the sentence, the model generates the target sentence word by word. Compared with previous methods such as RBMT and SMT, NMT does not need human-designed rules and features. NMT is an end-to-end framework that directly learns semantic representation and translation knowledge from the training corpora. With these advantages, NMT is now the dominant method in the MT community.

In this section, we first introduce NMT models and their key components, including basic recurrent neural network (RNN)-based models and their improvements, as well as the state-of-the-art NMT architecture, Transformer. Next, we describe multilingual translation and discuss methods such as back-translation and pivot-based translation for making full use of data, and methods such as multitask learning and universal models for improving NMT. We then introduce the latest progress in ST, including the cascaded model that pipelines automatic speech recognition (ASR), MT, and text-to-speech (TTS), and the end-to-end approach that directly models speech recognition and MT.

### 2.1. NMT model

A typical NMT model is built based on a standard RNN or its alternative [18,19]. Given a source sentence $x = \{x_1, x_2, ..., x_{T_x}\}$ (where $T_x$ is the length of $x$), the encoder RNN compresses $x$ into the hidden states $h = \{h_1, h_2, \ldots, h_{T_x}\}$ as follows:

$$h_t = g(h_{t-1}, x_t, \theta) \tag{1}$$

where $g(\cdot)$ is the activation function of the network; $h_t$ and $x_t$ are the hidden state and the source token at time $t$, respectively; $t$ is the time step; $\theta$ is a set of model parameters. In the basic model, the encoder takes the last hidden state $h_{T_x}$ as the representation of the source sentence. Then the decoder RNN produces translations as follows:

$$p(y|x) = \prod_{t=1}^{T_y} p(y_t | y_{<t}, \boldsymbol{c}) \tag{2}$$

where $y = \{y_1, y_2, \ldots, y_{T_y}\}$ is the target sentence, $p(y|x)$ is the translation probability, $T_y$ is the length of $y$, $\boldsymbol{c}$ is a vector generated from the hidden states $h$, $y_t$ is the target word, $y_{<t} = \{y_1, y_2, \ldots, y_{t-1}\}$ contains the target words that have been already generated.

One of the weaknesses of the standard RNN model is that the information decays rapidly during transmission in the network; thus, the translation quality drops heavily for long sentences. To overcome this issue, Bahdanau et al. [18] proposed three novel improvements, which are widely used in NMT models. These are described one by one below.

#### 2.1.1. The attention mechanism

When generating a target word, instead of using the last hidden state $h_{T_x}$ to represent the source sentence, the attention mechanism computes the association between the target token and all the source words, and evaluates how strong the association is.

$$\boldsymbol{c}_t = \sum_{j=1}^{T_x} a_{tj} h_j \tag{3}$$

where $\boldsymbol{c}_t$ is the contextual vector, $h_j$ is the hidden state of the source word $x_j$, $j$ is the word index of $x$, $a_{tj}$ is the association weight of the target word $y_t$ and $h_j$, which is computed as follows:

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_{i=1}^{T_x} \exp(e_{ti})} \tag{4}$$

where $e_{tj}$ is the alignment model parametrized as a feed-forward neural network, $i$ is the word index of $x$.

In fact, the attention mechanism is analogous to the "word alignment" used in SMT. Word alignment in SMT is a "hard alignment," which indicates whether a source word and a target word have a link or not, while the attention mechanism is a "soft alignment," which links a target word to all source words with different weights. The attention mechanism significantly improves the translation quality, making NMT a breakthrough technology in the MT history.

#### 2.1.2. Bidirectional encoding

Instead of a unidirectional encoder, some methods employ a bidirectional encoder. To be specific, a bidirectional encoder computes hidden states from both the left-to-right and right-to-left directions, such as $\overrightarrow{h} = \{\overrightarrow{h}_1, \overrightarrow{h}_2, \ldots, \overrightarrow{h}_{T_x}\}$ and $\overleftarrow{h} = \{\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_{T_x}\}$. The hidden states are then concatenated as $h = \left\{\left[\overrightarrow{h}_1, \overleftarrow{h}_1\right], \left[\overrightarrow{h}_2, \overleftarrow{h}_2\right], \cdots, \left[\overrightarrow{h}_{T_x}, \overleftarrow{h}_{T_x}\right]\right\}$. Thus, the hidden state contains both the history and the future information of the source sentence, which again improves the translation quality.

#### 2.1.3. Gated recurrent unit

The gated recurrent unit (GRU) is an alternative to conventional simple activation functions. GRU is analogous to long short-term memory (LSTM) [46], but is much more efficient. Both GRU and LSTM allow the network to learn long-distance dependency without suffering too much from the gradient vanishing problem [47].

Preliminary experiments on NMT showed significant improvements over conventional SMT. However, the early NMT models still had weaknesses, such as the out-of-vocabulary (OOV) problem, under-translation, and a slow decoding speed. To overcome these problems, He et al. [48] proposed the incorporation of statistical features such as the phrase table, the $n$-gram language model, and the length penalty into NMT. Along with this direction, researchers borrowed ideas from SMT and incorporated into NMT rich features, such as coverage [49], alignment agreement [50], syntax information [51–53], phrase tables [54,55], and translation recommendations [56]. Sennrich et al. [57] used the compression algorithm byte-pair-encoding [58] for word segmentation that compacts open vocabularies into a fixed-size vocabulary of subwords. This method is simple and efficient, so it is widely used in NMT for addressing the translation of OOV words and rare words.

Aside from RNN, researchers have put forward other model architectures. One weakness of RNN-based NMT is its lack of parallelization capability, as the computation of the current word depends on the previous words. Convolutional neural networks (CNNs), which are commonly used in computer vision, have been introduced to NMT [23]. Compared with RNN, a convolutional network creates hierarchical representations over sequences with short paths to capture long-distance dependencies, which makes the computations fully parallelized during training.

Inspired by the CNN NMT methods, Vaswani et al. [24] proposed a novel network named Transformer, which is built solely on attention mechanisms without any recurrences or convolutions. There are three kinds of attention in this method: encoder self-attention, decoder-masked attention, and encoder–decoder attention. The researchers proposed a novel scaled dot-product method to represent these kinds of attention.

$$\text{Attention}\ (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{5}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the query, key, and value vectors, respectively; $\sqrt{d}$ is the scaling factor; $\boldsymbol{K}^{\mathrm{T}}$ is the transpose of $\boldsymbol{K}$. More specifically, for each word, the model creates three vectors—a query vector, a key vector, and a value vector—by multiplying the word embeddings with different parameter metrics. The role of the attention is to compute a weighted sum of the values as an output that will be transferred to the next layer.

In addition, the researchers proposed a multihead attention mechanism.

$$\text{Multihead}\ (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}\ (\text{head}_1, \ldots, \text{head}_M)W^O \tag{6}$$

where $M$ is the head number, $\text{head}_m = \text{Attention}\ (\boldsymbol{Q}W_m^{\boldsymbol{Q}}, \boldsymbol{K}W_m^{\boldsymbol{K}}, \boldsymbol{V}W_m^{\boldsymbol{V}})(1 \le m \le M)$ denotes different attention spaces, and $W_m^{\boldsymbol{Q}}$, $W_m^{\boldsymbol{K}}$, $W_m^{\boldsymbol{V}}$, and $W^O$ are parameter matrices. The function $\text{Concat}(\text{head}_1, \ldots, \text{head}_M)$ concatenates all heads together.

Compared with recurrent and convolutional networks, Transformer has stronger parallelization and representation ability; thus, it achieves state-of-the-art performance not only in MT, but also in many other NLP tasks, such as the recent well-known pretraining models: bidirectional encoder representation from transformer (BERT) [59] and enhanced representation through knowledge integration (ERNIE) [60].

The above models are autoregressive models in which each output word depends on previous outputs. This setup limits the models' parallelization capability during decoding. Gu et al. [25] proposed a non-autoregressive transformer (NAT), which can generate target sequences in parallel.

$$p(y|x) = p_L(T|x;\phi) \cdot \prod_{t=1}^{T} p(y_t|x;\phi) \tag{7}$$

where $T$ is the length of the target sentence, which is modeled with a conditional distribution $p_L(T|x;\phi)$; $\phi$ is a set of model parameters.

Unlike autoregressive models, which stop decoding when generating the special token end-of-sentence (</s>), non-autoregressive models must first predict the length of the target sequence with $p_L(T|x;\phi)$. Although NAT achieves remarkable speedup during decoding, the translation quality is greatly degraded. The main possible reason is that it does not model word dependency, which is crucial for translation improvement. Encouraged by the decoding efficiency, researchers have proposed many methods to improve the non-autoregressive models, including knowledge distillation [61], imitation learning [26], and curriculum learning [62].

### 2.2. Multilingual translation

Different languages have different morphologies and structures, which makes translation among languages a difficult task—not only for MT, but also for human experts. For example, Chinese and English are subject-verb-object languages, while Japanese and Korean are subject-object-verb languages. When performing translation between Chinese and Japanese, long-distance reordering is usually required. In addition, Chinese is an isolating language with few morphological changes, while Japanese is an agglutinative language with rich word morphological changes. All of these differences make multilingual MT particularly difficult.

Data-driven MT methods—that is, SMT or NMT—attempt to learn the translation knowledge from a large quantity of parallel data. In general, an increased amount of training data results in an improved translation quality. Koehn and Knowles [63] showed that when the training words increased from 0.4 million to 385.7 million for English–Spanish translation, the translation quality improved by about 30% (absolute) in terms of bilingual evaluation understudy (BLEU) score.

Unfortunately, most of the world's languages lack parallel data, and are thus referred to as "resource-poor" languages. Building an NMT system for these languages is a great challenge due to the data sparseness problem. According to Internet World Stats, the number of users of the world's top ten languages (English, Chinese, Spanish, Arabic, Portuguese, Indonesian/Malay, French, Japanese, Russian, and German) on the Internet account for about 77% of the total number of Internet users. Of these, English and Chinese users account for 25.9% and 19.4%, respectively, while the sum of all other language users only accounts for 23.1%. For resource-rich languages such as Chinese and English, it is possible to collect billions of sentence pairs to train an MT model; however, for resource-poor language pairs such as Chinese–Hindi or Chinese–Kiswahili, there are only thousands of sentence pairs or less.

In addition, the deployment of multilingual translation systems costs a great deal. If we suppose that translation will be performed among $N$ languages ($N$ is the number of languages), it is usually necessary to build a translation model for each translation direction (e.g., Chinese-to-English and English-to-Chinese are two translation directions). In this case, it is necessary to build $N \times (N-1)$ translation models for $N$ languages.

With the success of NMT models, researchers have been seeking new ways to overcome the above challenges. In general, there are two kinds of methods for multilingual translation: methods that make full use of data and methods that improve NMT models.

Since multilingual translation among resource-poor languages lacks training data, it can be intuitively seen that it is necessary to collect more training data and make full use of the potential of that data. Compared with parallel corpus collection, it is easier to obtain a large amount of monolingual corpus. In NMT, the monolingual corpus is usually used for training data augmentation. One widely used method is back-translation [64,65], in which the main idea is to first train a standard NMT model on a small parallel corpus, and then use the model to translate a large quantity of monolingual data (e.g., sentences in the target languages) into the other side, so as to generate a "pseudo bilingual corpus" that can be used to retrain the translation model. In an extreme case, there may be no parallel corpus at all. To solve this problem, unsupervised translation methods have been proposed to build translation systems that are only based on the source and target monolingual corpora. Lample et al. [66] proposed mapping sentences in different languages into the same latent space and learning to translate by reconstructing sentences. Artetxe et al. [67] used an improved SMT model to initialize an unsupervised NMT model in order to further improve translation quality. Song et al. [29], Conneau and Lample [68], and Ren et al. [69] proposed an unsupervised NMT model to leverage the pretraining method.

Another research line is to leverage the resource-rich languages in order to improve the translation of resource-poor languages. This method can date back to the SMT era. The most widely used method is pivot-based translation, in which a high-resource language is used as the pivot language to build a bridge between low-resource language pairs [70]. For example, to develop a Chinese–German translation system, English can be chosen as the pivot language, since there is a large quantity of Chinese–English and English–German parallel data available. The simplest pivot-based translation method is the transfer method, which uses two cascaded translation systems [71,72]: the source–pivot translation system, which translates the source sentence into the pivot sentence; and the pivot–target translation system, which translates the pivot sentence into the target sentence. This method is widely used in practical systems because it is easy to implement. The weakness of this method is that the cascaded system suffers from the error propagation problem. Wu and Wang [73,74] and Cohn and Lapata [75] proposed a triangulation method to learn phrase-level translation knowledge by inducing a source–target translation model from source–pivot and pivot–target translation models.

NMT methods leverage the source-rich languages to improve the translation quality of resource-poor languages by using a universal model. Traditional MT methods require separate translation models for each language pair and each task, whereas NMT makes it possible to translate multiple languages across different tasks within a universal model. In general, this research can be classified into three categories: one-to-many, many-to-one, and many-to-many (M2M), depending on the number of languages on the source and target side.

Dong et al. [20] proposed a novel multitask learning method for multilingual NMT. As Fig. 1 shows, by sharing source representations with a shared-encoder, the model can make full use of the source language corpus across different language pairs. This method provides a unified framework for exploring the problem of translating one source language into multiple target languages. To deploy translation systems among $N$ languages, the model only needs to train $N$ encoders. Luong et al. [76] extended the framework to multitasks, including translation, parsing, and image captioning. Zoph and Knight [77] proposed a many-to-one NMT model that shares the decoder on the target side. Firat et al. [78] used different encoders and decoders with a shared attention mechanism for M2M translation.

Johnson et al. [79] proposed a simple approach that put all languages together to train a single encoder–decoder model to perform multilingual translation. The researchers added a special token to the beginning of the source sentence to indicate which target language it is translated into. This approach allows the NMT model to learn shared representations for linguistically similar languages [80], so no change is made to the NMT model architecture. Considering the diversity of languages, Tan et al. [81] studied how to group languages into several clusters and train a single NMT model for each cluster.

In a practical system, a hybrid translation method is usually used, which combines the above methods while considering translation efficiency, deployment cost, and so forth. Thanks to technological progress, current translation systems can support translation among hundreds of languages. Arivazhagan et al. [82] proposed a method for a massively multilingual MT that trains a single model with over 50 billion parameters on more than 25 billion sentence pairs, from 103 languages to and from English. Fan et al. [83] proposed an M2M-100 model that is trained on 7.5 billion sentence pairs and can perform translation between any pair of 100 languages.

### 2.3. Simultaneous translation

ST aims to achieve real-time translation with high quality and an as-short-as-possible delay between the source language speech and the translation output. In full-sentence translation (Section 2.1), each target word $y_t$ is predicted using the entire source sentence $x$. In ST, however, it is necessary to translate concurrently with the (growing) source sentence.

Research on ST falls into two categories: the cascaded (pipeline) method and the end-to-end method.

#### 2.3.1. Simultaneous S2S translation pipeline

A typical cascaded ST system consists of an ASR system that transcribes the source speech into source streaming text, an MT system that performs the translation from the source text into the target text and, finally, a TTS system to generate the target–language speech, as illustrated in Fig. 2. In practice, the TTS system is optional, depending on whether the output is text or speech in different application scenarios.

As mentioned, one of the biggest challenges in ST is achieving high translation quality with low latency. The streaming ASR output has no segmentation boundaries, while traditional MT systems take sentences with clear boundaries as input. Thus, there is a gap between the output of ASR and the input of MT. If a translation starts before adequate source content has been delivered, the translation quality degrades. However, waiting for too many source words increases the latency.

In general, two types of recent work split the ASR output into appropriate segments for the downstream MT system: fixed policies that consider fixed-length contexts and adaptive policies that obtain source segments dynamically.

Fixed policies are hard policies that follow a predefined schedule that is independent of the context. Such policies segment the source text based on a fixed length [43,84]. Ma et al. [43] proposed a simple wait-$k$ policy under a prefix-to-prefix architecture, where $k$ is the number of words that the model firstly read, and then

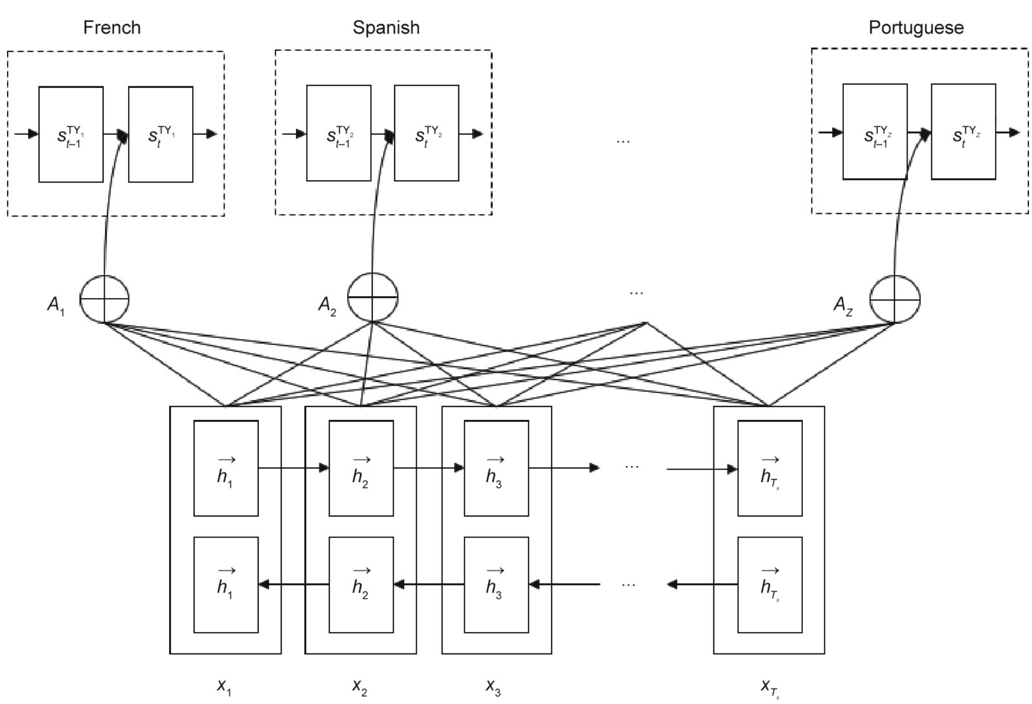


**Fig. 1.** Illustration of a multitask learning NMT framework for one-to-many translation. $A_1, A_2, \cdots, A_Z$ are the attentions for target languages; $\mathrm{TY}_1, \mathrm{TY}_2, \cdots, \mathrm{TY}_Z$ are target languages; $Z$ is the number of target languages; $s_t^{\mathrm{TY}_z} (1 \le z \le Z)$ are the hidden states on the decoding side.

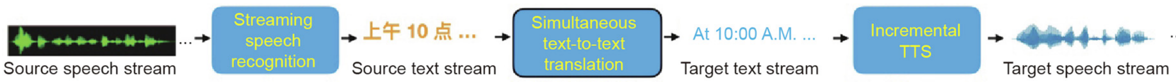


**Fig. 2.** Framework of the cascaded ST system.

translates concurrently with the rest of the source sentence; that is, the output is always $k$ words after the input. This policy was inspired by human simultaneous interpreters, who generally start translating a few seconds after the speaker's speech, and who finish with a few seconds' delay after the speaker finishes. For example, if $k = 2$, the first target word is predicted using the first two source words, the second target word is predicted using the first three source words and the generated target word, and so forth. More formally, Ma et al. [43] used the source prefix $\{x_1, x_2, \ldots, x_{q(t)}\}$ rather than the whole source sentence to predict $y_t : p(y_t|y_{<t}, x_{\leq q(t)})$, where $q(t)$ is a monotonic non-decreasing function that denotes the number of source words processed by the encoder when predicting $y_t$. Generally speaking, $q(t)$ can be used to represent arbitrary policies. There are two special cases where $q(t)$ is constant: ① $q(t) = |x|$, or full-sentence translation; and ② $q(t) = 0$, where $q(t)$ is an "oracle" that does not rely on any source information. It should be that, in any case, $0 \leq q(t) \leq |x|$ for all $t$. Policies of this type are simple and easy to implement. However, they do not dynamically consider suitable contextual information and usually result in a decrease in translation accuracy.

Adaptive policies learn to conduct source text segmentation according to dynamic contextual information. Such policies either use a specific model to chunk the streaming source text [85–89] or jointly learn segmentation and translation in an end-to-end framework [90,91]. Adaptive methods are more flexible than fixed ones, and achieve state-of-the-art translation results. Inspired by the chunking strategy used by human interpreters, Zhang et al. [92] proposed a novel method to detect meaningful units for ST. The streaming source text is dynamically split into segments that can be translated independently, which ensures the generation of a high-quality translation with low latency.

Incremental TTS, however, is a much less studied problem. Current state-of-the-art TTS systems generate speech after obtaining all the words in the texts, which results in high latency. In order to reduce latency, it is necessary to generate the speech incrementally. Conventional methods of incremental TTS are based on the Hidden Markov Model [93–97]. These models require full context labels of linguistic features, where each component is trained and tuned separately. Recent research has leveraged the strength of neural networks [98,99]. Yanagita et al. [98] proposed a segment-based TTS that synthesizes a segment at a time. Ma et al. [99] proposed a neural incremental word-level TTS. As shown in Fig. 3, this idea is based on two observations: ① The dependencies are very local; and ② audio playing is inherently sequential in nature, and can be done simultaneously with audio generation—that is, a segment of audio can be played while the subsequent text is being generated. To summarize, this method starts to generate a spectrogram for the first word after receiving the first two words; this spectrogram is fed into the vocoder to generate the waveform for the first word, which is also played immediately.

It is easy to implement a cascaded framework for ST. However, this framework suffers from several problems. For example, due to the simultaneity requirement, each of the three components should be simultaneous (streaming or incremental processing). Furthermore, the errors of each component propagate down the pipeline. A one-word error in the ASR system may make the translation result unacceptable. Thus, there is a need to develop more robust speech translation systems.

*2.3.2. Toward end-to-end ST*

The ultimate goal is to develop end-to-end ST technologies, so that the source language speech can be translated simultaneously into the target language without passing through intermediate stages, as in cascaded methods. This idea could not only reduce error propagation in the current pipeline, but also improve the efficiency of the system. However, it is extremely challenging to achieve both end-to-end translation and simultaneity together. Furthermore, the training data for an end-to-end ST model is very scarce. The available training data contains only hundreds of hours of speeches, most of which are for Japanese–English translation [100,101] and European languages [102,103]. For Chinese–English translation, Baidu has released an open dataset containing 70 h of speeches, including both the corresponding transcriptions and translations [104]. From the perspective of methodology, integrating speech recognition and translation into a unified framework is not trivial.

End-to-end ST is a cutting-edge technology. Bansal et al. [105] provides the first proof that end-to-end speech translation can be implemented without using any source transcriptions. Studies resort to pretraining or multitask learning to improve translation quality. Such studies either applied a pretrained encoder trained on ASR data [105], or leveraged the text translation to improve the speech translation [106–108]. Liu et al. [109] uses a knowledge distillation method to improve the end-to-end ST model by transferring the knowledge from the MT model. However, different tasks in these methods cannot share information with each other. To alleviate this problem, several studies proposed two-stage models [110–112], in which the decoder in the first stage performs recognition and generates a hidden state with which the second decoder conducts translation. Liu et al. [113] proposed an interactive end-to-end ST model that can conduct speech recognition and MT interactively, enhancing the performance on both tasks. Recent studies also tackle the issue of direct S2S translation [114,115]. However, due to the limited training data and the complexity of integrating speech recognition and MT into a uniform framework, the performance of current end-to-end ST methods does not yet meet the practical requirements.

At present, most practical ST systems use cascaded methods because they can be easily deployed and can generate high-quality translations. Xiong et al. [104] reported a comparison between a pipeline ST system and human interpreters with experience ranging from three to seven years. They found that the human interpreters usually skipped unimportant information to maintain a reasonable ear–voice span, which could result in a loss of adequacy but provided a shorter lag time, while the ST system produced more adequate translations. Shimizu et al. [100] shows that interpreters with less experience lose details during interpreting. These studies show that simultaneous interpreting remains a difficult task for both human interpreters and MT systems.

## 3. Applications of MT

MT is already widely used in many areas due to its low cost, high efficiency, and high translation quality. In China, typical human translation costs from 0.1 to 0.5 CNY per character, depending on the translator's proficiency, whereas MT systems cost about 0.00005 CNY per character. Fig. 4 shows the translation distribution among the top eight domains in Baidu Translate, which

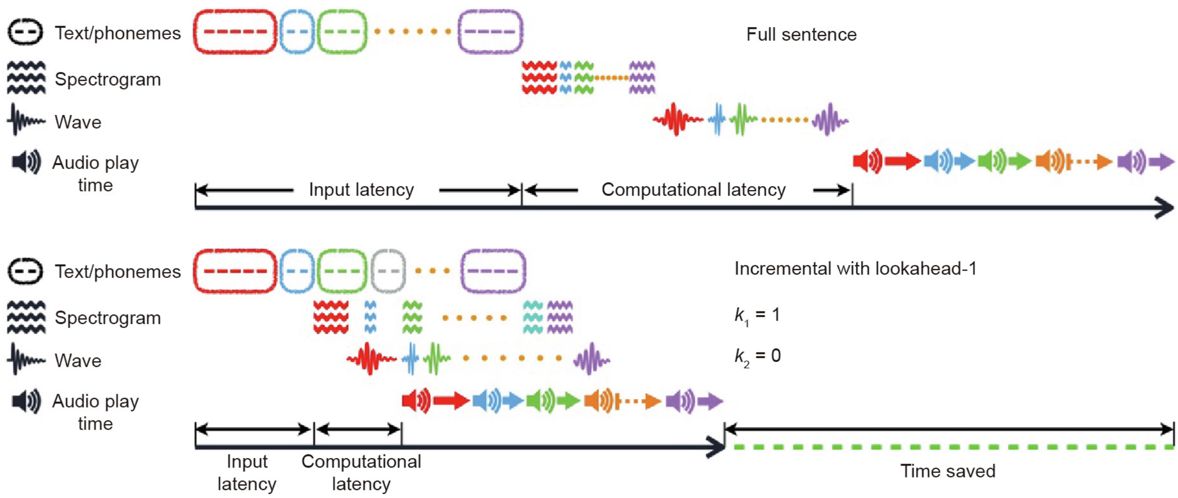


**Fig. 3.** Full-sentence TTS versus incremental TTS. $k_1$ and $k_2$ are the lookahead window sizes for spectrogram and wave generation, respectively.

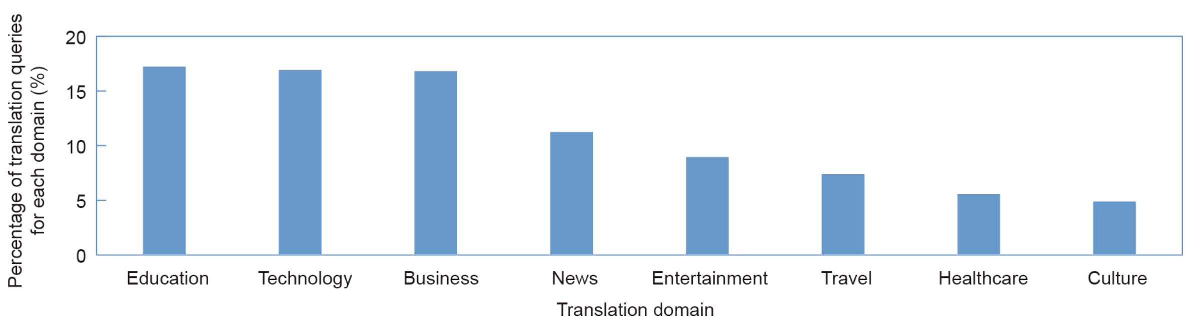


**Fig. 4.** Translation distribution of Baidu Translate.

supports translations between any pair among over 200 languages and supports translating queries with over 100 billion characters per day.

### 3.1. Text translation

Text translation is the most common form of MT application. Below are some typical applications of text translation.

(1) **Webpage translation.** With the rapid progress of globalization, there is an increasing need for quick acquisition of information in foreign languages. While it is expensive and time-consuming to hire human translators to translate a huge number of webpages, MT provides a convenient way to view webpages in foreign languages. Users just need to copy/paste the content of the webpage or input the uniform resource locator (URL) to read the pages in their own language.

(2) **Scientific literature translation.** Users such as researchers, engineers, and graduate students often use MT to read scientific literature such as papers and patents in their own language, or to translate their work into other languages. For example, translation in the domain of bio-medicine is growing rapidly in order to combat coronavirus disease 2019 (COVID-19). Scientific literature usually contains many terminologies. With domain adaptation technologies, a translation model can first be pretrained with a large training corpus, and then fine-tuned on a small amount of in-domain data for further improvement. In addition, formatted document translation is used to translate various kinds of documents, such as PowerPoint, Excel, Word, and portable document format (PDF), while keeping format information such as font size and font color.

(3) **E-commerce translation.** MT is widely used in transnational online trade. With the help of MT, sellers can effectively translate their website, product information, and manuals into foreign languages, while buyers can easily buy products from all over the world. MT is also used in customer services to improve service quality and efficiency.

(4) **Language learning.** Current MT systems usually provide rich functions, including translation, high-quality dictionaries, sentence pair examples, and so forth. Users can thus conveniently determine the meaning of a word or phrase and learn how to use it. Student users often input a whole paragraph for comprehension reading and use the sentence pair examples to help in their writing.

In addition to text translation, image translation and speech translation have been widely used in real applications based on recent advances in artificial intelligence techniques.

### 3.2. Image translation

Image translation combines computer vision and MT, as it takes images as input and then translates them into the target languages.

(1) **Multilingual image captioning.** This type of MT, which describes the content of pictures and performs visual question answering, has been widely studied in recent years [116–118]. Multilingual image translation borrows the idea of NMT, where the input of the encoder is an image and the output of the decoder is text. Since the model can generate different languages for the same picture, this function is very helpful for language studies.

(2) **Optical character recognition translation.** This form of MT first recognizes the characters in a picture and then performs

translation and renders it to replace the original source text. This function is useful for translating menus, street nameplates, product descriptions, and so forth, when traveling to foreign countries. With recent studies on jointly modeling the text and layout information of document images [119], MT can also be used to translate scanned documents while keeping the original format information.

### 3.3. Speech translation

Speech translation combines speech processing and MT; it takes voice in the source language as its input and generates text or voice in the target language as its output.

(1) **Simultaneous translation.** As mentioned in Section 2.3, great progress has recently been made in ST, enabling many kinds of products to provide ST services. Speech-to-text (S2T) translation projects both the ASR output and the translation scripts onto a single screen for the users' convenience. However, the limited space on a screen usually only makes it possible to display the scripts of one language pair. Thus, it is difficult to extend S2T to multiple languages. S2S translation solves this problem by allowing the audience to listen to the target voice via their cell phones. Thus, users from different countries can choose to listen to their mother language or to whatever other language they prefer. ST systems are currently widely used in international conferences. Due to the COVID-19 pandemic, many more conferences are being held virtually—that is online. ST has been integrated into online meeting systems to provide real-time translation. In addition, users can use ST plugins to watch foreign videos, such as films and lectures, in their own language.

(2) **Portable translation devices.** These devices are capable of voice translation and have been widely favored by users in recent years. They are easy to carry and use in many scenarios, including language learning, overseas traveling, and business negotiation.

MT can also be used for poem generation [120] and Chinese couplet generation. Taking the former generated line as the "source" and the subsequent line as the "target," MT models can generate poems in a line-by-line manner.

## 4. Challenges and future directions

Although great progress has been achieved in MT, there is always room for improvement. At meetings such as Workshop on Statistical Machine Translation, it is sometimes suggested that machines are better than human translators. Certain metrics (i.e., BLEU, word error rate (WER), metric for evaluation of translation with explicit ordering (METEOR)) [121–123] and benchmarks may suggest that this is the case, but such metrics may not be measuring what is important. A good translation should have at least two characteristics: adequacy and fluency. Nowadays, NMT methods can produce translations for some language pairs and domains with very high adequacy and fluency in particular text translation scenarios; however, such methods are far from perfect, especially in ST scenarios. Many aspects remain to be improved.

First, new evaluation metrics are needed to evaluate what really matters. For example, human interpreters do not attempt to translate everything when performing simultaneous interpreting. It is important to know what needs to be said and when it needs to be said. Human interpreters know when they need to speed up and when they can take their time. They know what needs to be emphasized and what can be omitted. However, MT systems translate everything and do not know how to omit unimportant parts to reduce latency. Furthermore, emphasis is important in translation; a translation should reflect the emphasis that is present in the source. Recently, studies have investigated the use of acoustic clues to identify emphasis and translate it into the target language [124–126]. Besides speech information, the speaker's body language (and prosody) make it clear when the speaker is emphasizing a particular point (as opposed to a different point); nevertheless, it is difficult to synchronize the translation with the speaker's body language. Speakers often make references to slides; but again, it is difficult to synchronize a translation with reference to slides. Although metrics such as BLEU and WER reward completeness, many other aspects contribute to a good translation: latency, emphasis, synchronization, comprehension, and so forth. None of these metrics reward these aspects. The front-end ASR system should capture not only words, but also emphasis that would have consequences for downstream steps including translation and speech synthesis. Metrics need to be developed that can reward systems that emphasize what needs to be emphasized, while penalizing systems that translate trivial parts that should not be translated.

Second, the robustness of MT needs further improvement. Sometimes, a slight change in the source sentence—such as a word or punctuation mark—can lead to great changes in the translation. However, human beings have a strong error-tolerant ability that allows them to flexibly deal with various non-standard language phenomena and errors, and sometimes even unconsciously correct them. Robust MT systems are crucial in real applications. Developing explainable MT methods may be one possible solution.

Third, NMT methods are facing serious data sparseness problems in resource-poor language pairs and domains. The current MT systems often use tens of millions or even hundreds of millions of sentence pairs of data for training. Otherwise, the translation quality will be poor. However, human beings can learn from only a small number of samples. Although many data-augmentation methods, multitask learning methods, and pretraining methods have been proposed to alleviate this problem, the question of how to improve the translation quality for resource-poor language pairs remains open.

In summary, there is still a long way to go to achieve high-quality MT. It is necessary to develop new methods that can combine symbolic rules, knowledge, and neural networks to further improve translation quality. Fortunately, the use of MT in real applications continues to provide more data, promoting the quick development of new MT methods.

## Compliance with ethics guidelines

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church declare that they have no conflict of interest or financial conflicts to disclose.

## References

[1] Weaver W. Translation. Mach Transl Lang 1955;14:15–23.
[2] Hutchins J. ALPAC: the (in) famous report. In: Nirenburg S, Somers HL, Wilks YA, editors. Readings in machine translation. Cambridge: MIT Press; 2003.
[3] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle. In: Elithorn A, Banerji R, editors. In: Proceedings of the International NATO Symposium on Artificial and Human Intelligence. New York City: Elsevier North-Holland, Inc; 1984. p. 173–80.
[4] Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty JD, et al. A statistical approach to machine translation. Comput Linguist 1990;16(2):79–85.
[5] Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL. The mathematics of statistical machine translation: parameter estimation. Comput Linguist 1993;19(2):263–311.
[6] Church KW, Mercer RL. Introduction to the special issue on computational linguistics using large corpora. Comput Linguist 1993;19(1):1–24.
[7] Al-Onaizan Y, Curin J, Jahr M, Knight K, Lafferty J, Melamed D, et al. Statistical machine translation: final report. Baltimore: Johns Hopkins University Summer Workshop; 1999.
[8] Och FJ, Ney H. A systematic comparison of various statistical alignment models. Comput Linguist 2003;29(1):19–51.
[9] Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proceedings of the Human Language Technology Conference of the North American

Chapter of the Association for Computational Linguistics; 2003 May 27–Jun 1; Edmonton, AB, Canada; 2003.
[10] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, et al. Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.
[11] Wang H. [Multi-strategy machine translation]. In: Cao YQ, Sun MS, editors. [Frontiers of Chinese information processing]. Beijing: Tsinghua University Press; 2006. p. 45–52. Chinese.
[12] Koehn P, Hoang H. Factored translation models. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CoNLL); 2007 Jun 25–27; Prague, Czech Republic; 2007.
[13] Chiang D. Hierarchical phrase-based translation. Comput Linguist 2007;33(2): 201–28.
[14] Yamada K, Knight K. A syntax-based statistical translation model. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics; 2001 Jul 6–11; Toulouse, France; 2001.
[15] Galley M, Graehl J, Knight K, Marcu D, DeNeefe S, Wang W, et al. Scalable inference and training of context-rich syntactic translation models. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics; 2006 Jul 17–21; Sydney, NSW, Australia; 2006.
[16] Liu Y, Liu Q, Lin S. Tree-to-string alignment template for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics; 2006 Jul 17–21; Sydney, NSW, Australia; 2006.
[17] Graehl J, Knight K, May J. Training tree transducers. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; 2004 May 2–7; Boston, MA, USA; 2004.
[18] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, USA; 2015.
[19] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems; 2014 Dec 8–13; Montreal, QC, Canada; 2014.
[20] Dong D, Wu H, He W, Yu D, Wang H. Multi-task learning for multiple language translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015 Jul 26–31; Beijing, China; 2015.
[21] Pouliquen B. WIPO Translate: patent neural machine translation publicly available in 10 languages [presentation]. In: Machine Translation XVI; 2017 Sep 18–22; Nagoya, Japan; 2017.
[22] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. arXiv: 1609.08144.
[23] Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia; 2017.
[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA; 2017.
[25] Gu J, Bradbury J, Xiong C, Li VOK, Socher R. Non-autoregressive neural machine translation. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.
[26] Wei B, Wang M, Zhou H, Lin J, Xie J, Sun X. Imitation learning for non-autoregressive neural machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.
[27] Lample G, Conneau A, Denoyer L, Ranzato M. Unsupervised machine translation using monolingual corpora only. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.
[28] Artetxe M, Labaka G, Agirre E. An effective approach to unsupervised machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.
[29] Song K, Tan X, Qin T, Lu J, Liu TY. Mass: masked sequence to sequence pre-training for language generation. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA; 2019.
[30] Kato Y. The future of voice-processing technology in the world of computers and communications. Pro Natl Acad Sci USA 1995;92(22):10060–3.
[31] Tomita M, Tomabechi H, Saito H. SpeechTrans: an experimental real-time speech-to-speech translation. Lang Res 1990;26(4):663–72.
[32] Kitano H. Speech-to-speech translation: a massively parallel memory-based approach. Boston: Kluwer Academic Publishers; 1994.
[33] Waibel A, Jain AN, McNair AE, Saito H, Hauptmann AG, Tebelskis J. JANUS: a speech-to-speech translation using connectionist and symbolic processing strategies. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing; 1991 Apr 14–17; Toronto, ON, Canada; 1991.
[34] Morimoto T, Takezawa T, Yato F, Sagayama S, Tashiro T, Nagata M, et al. ATR's speech translation system: ASURA. In: Proceedings of the 3rd European Conference on Speech Communication and Technology; 1993 Sep 22–25; Berlin, Germany; 1993.
[35] Roe DB, Pereira FCN, Sproat RW, Riley MD, Moreno PJ, Macarron A. Efficient grammar processing for a spoken language translation system. In: Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing; 1992 Mar 23–26; San Francisco, CA, USA; 1992.
[36] Sumita E, Shimizu T, Nakamura S. NICT-ATR speech-to-speech translation system. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.
[37] Fügen C, Kolss M, Paulik M, Stüker S, Schultz T, Waibel A. Open domain speech translation: from seminars and speeches to lectures. In: Proceedings of TC-STAR Workshop on Speech-to-Speech Translation; 2006 Jun 19–21; Barcelona, Spain; 2006.
[38] Moser-Mercer B, Künzli A, Korac M. Prolonged turns in interpreting: effects on quality, physiological and psychological stress (pilot study). Interpreting 1998;3(1):47–64.
[39] Wang H, Wu H, Hu X, Liu Z, Li J, Ren D, et al. The TCH machine translation system for IWSLT 2008. In: Proceedings of International Workshop on Spoken Language Translation; 2008 Oct 20–21; Honolulu, HI, USA; 2008.
[40] Nakamura S, Markov K, Nakaiwa H, Kikui G, Kawai H, Jitsuhiro T, et al. The ATR multilingual speech-to-speech translation system. IEEE Trans Audio Speech Lang Process 2006;14(2):365–76.
[41] He H, Boyd-Graber J, Daume H III. Interpretese vs. translationese: the uniqueness of human strategies in simultaneous interpretation. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA; 2016.
[42] Wang H, Gao W, Li S. Utterance segmentation of spoken Chinese. Chin J Comput 1999;22(10):1009–13. Chinese.
[43] Ma M, Huang L, Xiong H, Zheng R, Liu K, Zhang B, et al. STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.
[44] Zhang JJ, Zong CQ. Neural machine translation: challenges, progress and future. Sci China Technol Sci 2020;63(10):2028–50.
[45] Edunov S, Ott M, Auli M, Grangier D. Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31–Nov 4; Brussels, Belgium; 2018.
[46] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[47] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 1994;5(2):157–66.
[48] He W, He Z, Wu H, Wang H. Improved neural machine translation with SMT features. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence; 2016 Feb 12–17; Phoenix, AZ, USA; 2016.
[49] Tu Z, Lu Z, Liu Y, Liu X, Li H. Modeling coverage for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany; 2016.
[50] Cheng Y, Shen S, He Z, He W, Wu H, Sun M, et al. Agreement-based joint training for bidirectional attention-based neural machine translation. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence; 2016 Jul 9–15; New York City, NY, USA; 2016.
[51] Sennrich R, Haddow B. Linguistic input features improve neural machine translation. In: Proceedings of the First Conference on Machine Translation; 2016 Aug 11–12; Berlin, Germany; 2016.
[52] Wu S, Zhou M, Zhang D. Improved neural machine translation with source syntax. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017 Aug 19–25; Melbourne, VIC, Australia; 2017.
[53] Li J, Xiong D, Tu Z, Zhu M, Zhang M, Zhou G. Modeling source syntax for neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017 Jul 30–Aug 4; Vancouver, CB, Canada; 2017.
[54] Feng Y, Zhang S, Zhang A, Wang D, Abel A. Memory-augmented neural machine translation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017 Sep 7–11; Copenhagen, Denmark; 2017.
[55] Zhao Y, Wang Y, Zhang J, Zong C. Phrase table as recommendation memory for neural machine translation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden; 2018.
[56] Wang X, Lu Z, Tu Z, Li H, Xiong D, Zhang M. Neural machine translation advised by statistical machine translation. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017 Feb 4–9; San Francisco, CA, USA; 2017.
[57] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany; 2016.
[58] Gage P. A new algorithm for data compression. C Users J 1994;12(2):23–38.
[59] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minnesota, MN, USA; 2019.
[60] Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, et al. ERNIE 2.0: a continual pre-training framework for language understanding. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA; 2020.

[61] Zhou C, Neubig G, Gu J. Understanding knowledge distillation in non-autoregressive machine translation. 2019. arXiv:1911.02727.
[62] Guo J, Tan X, Xu L, Qin T, Chen E, Liu TY. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA; 2020.
[63] Koehn P, Knowles R. Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation; 2017 Aug 4; Vancouver, CB, Canada; 2017.
[64] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany; 2016.
[65] Poncelas A, Shterionov D, Way A, Wenniger GMDB, Passban P. Investigating backtranslation in neural machine translation. 2018. arXiv:1804.06189.
[66] Lample G, Conneau A, Denoyer L, Ranzato M. Unsupervised machine translation using monolingual corpora only. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.
[67] Artetxe M, Labaka G, Agirre E. An effective approach to unsupervised machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.
[68] Conneau A, Lample G. Cross-lingual language model pretraining. In: Proceedings of the 33rd Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada; 2019.
[69] Ren S, Wu Y, Liu S, Zhou M, Ma S. Explicit cross-lingual pre-training for unsupervised machine translation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China; 2019.
[70] Wang H, Wu H, Liu Z. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In: Proceedings of the COLING/ACL2006 Main Conference Poster Sessions; 2006 Jul 17–21; Sydney, NSW, Australia; 2006.
[71] Utiyama M, Isahara H. A comparison of pivot methods for phrase-based statistical machine translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics; 2007 Apr 22–27; Rochester, NY, USA; 2007.
[72] Khalilov M, Costa-Jussà MR, Henríquez CA, Fonollosa JAR, Hernández A, Mariño JB, et al. The TALP & I2R SMT systems for IWSLT 2008. In: Proceedings of the International Workshop on Spoken Language Translation; 2008 Oct 20–21; Honolulu, HI, USA; 2008.
[73] Wu H, Wang H. Pivot language approach for phrase-based statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.
[74] Wu H, Wang H. Revisiting pivot language approach for machine translation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language; 2009 Aug 2–7; Singapore; 2009.
[75] Cohn T, Lapata M. Machine translation by triangulation: making effective use of multi-parallel corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics; 2007 Jun 25–27; Prague, Czech Republic; 2007.
[76] Luong MT, Le QV, Sutskever I, Vinyals O, Kaiser L. Multi-task sequence to sequence learning. In: Proceedings of the International Conference on Learning Representations; 2016 May 2–4; San Juan, Puerto Rico; 2016.
[77] Zoph B, Knight K. Multi-source neural translation. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA; 2016.
[78] Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA; 2016.
[79] Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. Trans Assoc Comput Linguist 2017;5:339–51.
[80] Kudugunta S, Bapna A, Caswell I, Firat O. Investigating multilingual NMT representations at scale. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China; 2019.
[81] Tan X, Chen J, He D, Xia Y, Qin T, Liu TY. Multilingual neural machine translation with language clustering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China; 2019.
[82] Arivazhagan N, Bapna A, Firat O, Lepikhin D, Johnson M, Krikun M, et al. Massively multilingual neural machine translation in the wild: findings and challenges. 2019. arXiv:1907.05019.
[83] Fan A, Bhosale S, Schwenk H, Ma Z, El-Kishky A, Goyal S, et al. Beyond English-centric multilingual machine translation. 2020. arXiv:2010.11125.
[84] Dalvi F, Durrani N, Sajjad H, Vogel S. Incremental decoding and training methods for simultaneous translation in neural machine translation. 2018. arXiv:1806.03661.
[85] Sridhar VKR, Chen J, Bangalore S, Ljolje A, Chengalvarayan R. Segmentation strategies for streaming speech translation. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2013 Jun 9–14; Atlanta,GA; USA; 2013.
[86] Oda Y, Neubig G, Sakti S, Toda T, Nakamura S. Optimizing segmentation strategies for simultaneous speech translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014 Jun 23–25; Baltimore, MD, USA; 2014.
[87] Cho K, Esipova M. Can neural machine translation do simultaneous translation? 2016. arXiv:1606.02012.
[88] Gu J, Neubig G, Cho K, Li VOK. Learning to translate in real-time with neural machine translation. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; 2017 Apr 3–7; Valencia, Spain; 2017.
[89] Fujita T, Neubig G, Sakti S, Toda T, Nakamura S. Simple, lexicalized choice of translation timing for simultaneous speech translation. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association; 2013 Aug 25–29; Lyon, France; 2013.
[90] Arivazhagan N, Cherry C, Macherey W, Chiu CC, Yavuz S, Pang R, et al. Monotonic infinite lookback attention for simultaneous machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy; 2019.
[91] Ma X, Pino J, Cross J, Puzon L, Gu J. Monotonic multihead attention. In: Proceedings of the International Conference on Learning Representations; 2020 Apr 26–May 1; Addis Ababa, Ethiopia; 2020.
[92] Zhang R, Zhang C, He Z, Wu H, Wang H. Learning adaptive segmentation policy for simultaneous translation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; 2020 Nov 16–20; online; 2020.
[93] Baumann T. Partial representations improve the prosody of incremental speech synthesis. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association; 2014 Sep 14–18; Singapore; 2014.
[94] Baumann T. Decision tree usage for incremental parametric speech synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; 2014 May 4–9; Florence, Italy; 2014.
[95] Pouget M, Hueber T, Bailly G, Baumann T. HMM training strategy for incremental speech synthesis. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association; 2015 Sep 6–10; Dresden, Germany; 2015.
[96] Pouget M, Nahorna O, Hueber T, Bailly G. Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis. In: Proceedings of Inter-Speech 2016; 2016 Sep 8–12; San Francisco, CA, USA; 2016.
[97] Yanagita T, Sakti S, Nakamura S. Incremental TTS for Japanese language. In: Proceedings of Interspeech; 2018 Sep 2–6; Hyderabad, India; 2018.
[98] Yanagita T, Sakti S, Nakamura S. Neural iTTS: toward synthesizing speech in real-time with end-to-end neural text-to-speech framework. In: Proceedings of the 10th ISCA Speech Synthesis Workshop; 2019 Sep 20–22; Vienna, Austria; 2019.
[99] Ma M, Zheng B, Liu K, Zheng R, Liu H, Peng K, et al. Incremental text-to-speech synthesis with prefix-to-prefix framework. In: Findings of the Association for Computational Linguistics: EMNLP 2020; 2020 Nov 16–20; online; 2020.
[100] Shimizu H, Neubig G, Sakti S, Toda T, Nakamura S. Collection of a simultaneous translation corpus for comparative analysis. In: Proceedings of Ninth International Conference on Language Resources and Evaluation; 2014 May 26–31; Reykjavik, Iceland; 2014.
[101] Toyama H, Ryu K, Matsubara S, Kawaguchi, Nobuo K, Inagaki Y. CIAIR simultaneous interpretation corpus. In: Proceedings of Oriental COCOSDA; 2004 Nov 17–19; New Delhi, India; 2004.
[102] Sandrelli A, Bendazzoli C. Tagging a corpus of interpreted speeches: the European parliament interpreting corpus (EPIC). In: Proceedings of LREC; 2006 May 22–28; Genoa, Italy; 2004.
[103] Di Gangi MA, Cattoni R, Bentivogli L, Negri M, Turchi M. MuST-C: a multilingual speech translation corpus. In: Proceedings of 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minnesota, MN, USA; 2019.
[104] Xiong H, Zhang R, Zhang C, He Z, Wu H, Wang H. DuTongChuan: context–aware translation model for simultaneous interpreting. 2019. arXiv:1907.12984.
[105] Bansal S, Kamper H, Livescu K, Lopez A, Goldwater S. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics; 2018 Jun 1–6; New Orleans, LA, USA; 2018.
[106] Weiss RJ, Chorowski J, Jaitly N, Wu Y, Chen Z. Sequence-to-sequence models can directly translate foreign speech. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association; 2017 Aug 20–24; Stockholm, Sweden; 2017.
[107] Anastasopoulos A, Chiang D. Leveraging translations for speech transcription in low-resource settings. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association; 2018 Sep 2–6; Hyderabad, India; 2018.

[108] Bérard A, Pietquin O, Servan C, Besacier L, Servan C. Listen and translate: a proof of concept for end-to-end speech-to-text translation. In: Proceedings of the 30th Conference on Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain; 2016.

[109] Liu Y, Xiong H, Zhang J, He Z, Wu H, Wang H, et al. End-to-end speech translation with knowledge distillation. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association; 2019 Sep 15–19; Graz, Austria; 2019.

[110] Kano T, Sakti S, Nakamura S. Structured based curriculum learning for end-to-end English–Japanese speech translation. In: Proceedings of the 18th Annul Conference of the International Speech Communication Association; 2017 Aug 20–24; Stockholm, Sweden; 2017.

[111] Anastasopoulos A, Chiang D. Tied multitask learning for neural speech translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018 Jun 1–6; New Orleans, Louisiana; 2018.

[112] Sperber M, Neubig G, Niehues J, Waibel A. Attention-passing models for robust and data-efficient end-to-end speech translation. Transl Assoc Comput Linguist 2019;7:313–25.

[113] Liu Y, Zhang J, Xiong H, Zhou L, He Z, Wu H, et al. Synchronous speech recognition and speech-to-text translation with interactive decoding. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA; 2020.

[114] Jia Y, Weiss RJ, Biadsy F, Macherey W, Johnson M, Chen Z, et al. Direct speech-to-speech translation with a sequence-to-sequence model. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019); 2019 Sep 15–19; Graz, Austria; 2019.

[115] Kano T, Sakti S, Nakamura S. Transformer-based direct speech-to-speech translation with transcoder. In: Proceedings of the IEEE Spoken Language Technology Workshop; 2021 Jan 19–22; Shenzhen, China; 2021.

[116] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 8–10; Boston, MA, USA; 2015.

[117] Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA; 2017.

[118] Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom–up and top–down attention for image captioning and visual question answering. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018.

[119] Xu Y, Li M, Cui L, Huang S, Wei F, Zhou M. LayoutLM: pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Minning; 2020 Aug 23–27; online; 2020.

[120] Wang Z, He W, Wu H, Wu H, Li W, Wang H, et al. Chinese poetry generation with planning based neural network. Proceedings of the 26th International Conference on Computational Linguistics; 2016 Dec 11–16; Osaka, Japan; 2016.

[121] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul 7–12; Philadelphia, PA, USA; 2002.

[122] Tomás J, Mas JÀ, Casacuberta F. A quantitative method for machine translation evaluation. In: Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable? 2003 Apr 12–17; Budapest, Hungary; 2003.

[123] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; 2005 Jun 29; Ann Arbor, MI, USA; 2005.

[124] Tsiartas A, Georgiou PG, Narayanan SS. Toward transfer of acoustic cues of emphasis across languages. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association; 2013 Aug 25–29; Lyon, France; 2013.

[125] Do QT, Sakti S, Nakamura S. Sequence-to-sequence models for emphasis speech translation. IEEE/ACM Trans Audio Speech Lang Process 2018;26(10):1873–83.

[126] Do QT, Toda T, Neubig G, Sakti S, Nakamura S. Preserving word-level emphasis in speech-to-speech translation. IEEE/ACM Trans Audio Speech Lang Process 2017;25(3):544–56.