# A Survey on Low-Resource Neural Machine Translation

**Rui Wang**, **Xu Tan**, **Renqian Luo**, **Tao Qin** and **Tie-Yan Liu**
Microsoft Research Asia
{ruiwa, xuta, t-reluo, taoqin, tyliu}@microsoft.com

## Abstract

Neural approaches have achieved state-of-the-art accuracy on machine translation but suffer from the high cost of collecting large scale parallel data. Thus, a lot of research has been conducted for neural machine translation (NMT) with very limited parallel data, i.e., the low-resource setting. In this paper, we provide a survey for low-resource NMT and classify related works into three categories according to the auxiliary data they used: (1) exploiting monolingual data of source and/or target languages, (2) exploiting data from auxiliary languages, and (3) exploiting multi-modal data. We hope that our survey can help researchers to better understand this field and inspire them to design better algorithms, and help industry practitioners to choose appropriate algorithms for their applications.

## 1 Introduction

Powered by deep learning, neural machine translation (NMT) [Bahdanau *et al.*, 2014; Vaswani *et al.*, 2017] has become the dominant approach for machine translation. One limitation of NMT is that it needs a large scale of parallel data for model training. There are thousands of languages in the world[1] and unfortunately, most of them lack parallel data. Thus, popular commercial translators (e.g., Google translator, Microsoft Bing translator, Amazon translator) only support tens or a hundred languages. NMT has attracted much research attention for low-resource languages. Given that many models/algorithms have been proposed in recent years, a review on low-resource NMT is very helpful for fresh researchers entering this area and industry practitioners. Although there already exists surveys on many aspects of NMT (e.g., domain adaptation [Chu and Wang, 2018], multilingual translation [Dabre *et al.*, 2020]), a comprehensive survey for low-resource NMT is still lacking. Therefore, in this paper, we conduct a comprehensive and well-structured survey on low-resource NMT to fill in this blank.

**NMT basics.** An NMT model $\theta$ translates a sentence $\mathbf{x}$ in the source language to a sentence $\mathbf{y}$ in the target language. With a parallel training corpus $\mathbf{C}$, the model $\theta$ is trained by minimizing the negative log-likelihood loss: $L_\theta = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{C}} -\log P\left(\mathbf{y}|\mathbf{x};\theta\right)$. The encoder-decoder structure is widely used in NMT, where the encoder converts the source sentence into a sequence of hidden representations and the decoder generates target words conditioned on the source hidden representations and previously generated target words. The encoder and decoder can be recurrent neural networks [Dong *et al.*, 2015], convolutional neural networks [Gehring *et al.*, 2017], and Transformer [Vaswani *et al.*, 2017].

**Organization of this survey.** Due to the lack of parallel sentence pairs, leveraging data other than parallel sentences is essential in low-resource NMT. In this paper, we categorize existing algorithms on low-resource NMT into three categories according to the data they use to help a low-resource language pair:

- *Monolingual data.* Leveraging unlabeled data to boost machine learning models is a popular and effective approach in various areas. Similarly, in NMT, leveraging unlabeled monolingual data attracts lots of attentions (see Section 2) since collecting monolingual data is much easier and of lower cost than parallel data.
- *Data from auxiliary languages.* Languages with similar syntax and/or semantics are helpful to each other when training NMT models. Leveraging data from related and rich-resource languages has shown great success in low-resource NMT (see Section 3).
- *Multi-modal data.* Multi-modal data (e.g., parallel data between text and image) has also been used in low-resource NMT, as reviewed in Section 4.

In addition to reviewing algorithms, we also summarize widely used data corpora for low-resource NMT in Section 5. At last, we conclude this survey and point out future research directions in Section 6.

## 2 Exploiting Monolingual Data

Monolingual data contains a wealth of linguistic information (e.g., grammar and contextual information) and is more abundant and easier to obtain than bilingual parallel data, which is

[1] https://en.wikipedia.org/wiki/Language

useful to improve the translation quality especially in low-resource scenario. Plenty of works have exploited monolingual data in NMT systems, which we categorize into several aspects: (1) back translation, which is a simple and promising approach to take advantage of the target-side monolingual data [Sennrich *et al.*, 2016], (2) forward translation also called knowledge distillation, which utilizes the source-side monolingual data [Zhang and Zong, 2016b], (3) joint training on both translation directions, which can take advantage of the monolingual data on both the source and target sides [He *et al.*, 2016; Hoang *et al.*, 2018; Niu *et al.*, 2018; Zheng *et al.*, 2020], (4) unsupervised NMT, which builds NMT models with only monolingual data, and can be applied to the language pairs without any parallel data [Lample *et al.*, 2018a; Artetxe *et al.*, 2018], (5) pre-training, which leverages monolingual data with self-supervised training for language understanding and generation, and thus improves the quality of NMT models [Conneau and Lample, 2019; Song *et al.*, 2019; Lewis *et al.*, 2020], (6) comparable monolingual corpus, which contains implicit parallel information and can improve the translation quality [Wu *et al.*, 2019a], and (7) enhancing with bilingual dictionary, where the bilingual dictionary is used together with monolingual data to enhance the translation on low-resource languages. In this section, we provide an overview of these methods on exploiting monolingual data in NMT.

### 2.1 Back & Forward Translation

In back translation, pseudo parallel sentence pairs are generated by translating the target-side monolingual sentences to the source language via a translation system in the reverse direction [Sennrich *et al.*, 2016], while in forward translation, pseudo parallel sentence pairs are generated by translating the source-side monolingual sentences to the target language via a translation system in the same direction [Zhang and Zong, 2016b]. Then, the pseudo parallel data is mixed with the original parallel data to train an NMT model. It has been shown that back and forward translation provides promising performance gain on NMT systems [Sennrich *et al.*, 2016; Zhang and Zong, 2016b].

Besides the typically used beam search [Sennrich *et al.*, 2016; Zhang and Zong, 2016b], there are also some other methods to generate the pseudo parallel data: (1) random sampling according to the output probability distribution [Imamura *et al.*, 2018], (2) adding noise to source sentences generated by beam search [Edunov *et al.*, 2018], and (3) prepending a tag to the source sentences generated by beam search [Caswell *et al.*, 2019]. It is observed that random sampling and adding noise only works well on high resource setting compared to standard beam search [Edunov *et al.*, 2018], while prepending a tag performs the best on both high and low resource settings [Caswell *et al.*, 2019].

### 2.2 Joint Training on Both Translation Directions

Considering that both the source and target sides monolingual data has valuable information, some works leverage both of them via joint training on the two translation directions. Dual learning [He *et al.*, 2016; Qin, 2020] simultaneously improves the two models on both translation directions by aligning the original monolingual sentences $x$ and the sentences $x'$ translated forward and then backward ($x \to y' \to x'$) by the two models. Wang *et al.* [2019a] further improve dual learning by introducing multi-agent for both translation directions. Iterative back translation [Hoang *et al.*, 2018] and data diversification [Nguyen *et al.*, 2019] simultaneously trains one or multiple NMT models on each translation direction and iteratively updates the back-translated and forward-translated corpus via the updated better NMT models. Data Diversification [Nguyen *et al.*, 2019]. Bi-directional NMT [Niu *et al.*, 2018] trains both the translation directions in the same model with a tag indicating the direction at the beginning of source sentences, and then leverages both source-side and target-side monolingual data by back and forward translation. Mirror-generative NMT [Zheng *et al.*, 2020] jointly trains the translation models on both directions and the language models for both source and target languages with a shared latent variable.

### 2.3 Unsupervised NMT

To deal with the zero-resource translation scenario without any parallel sentences, a common approach is unsupervised learning for NMT [Lample *et al.*, 2018a; Artetxe *et al.*, 2018], which typically relies on two components to ensure the learning efficiency and quality: (1) bilingual alignment, which enables the model with good alignments between the two languages, and (2) translation improvement, which gradually improves the translation quality by iterative learning, typically through back translation [Sennrich *et al.*, 2016].

**Bilingual alignment.** How to initially align between the two languages is an open problem. There are mainly four kinds of approaches: (1) bilingual word embedding [Mikolov *et al.*, 2013], where the NMT system can either start from a word-by-word translation derived from the bilingual word embedding [Lample *et al.*, 2018a] or initialize the embedding parameters according to bilingual word embedding [Artetxe *et al.*, 2018; Yang *et al.*, 2018], (2) denoising auto-encoder (DAE) [Vincent *et al.*, 2008], which can build a shared latent space of two languages by learning to reconstruct sentences in the two languages from a noised version [Lample *et al.*, 2018a; Artetxe *et al.*, 2018; Yang *et al.*, 2018], (3) unsupervised statistical machine translation (SMT), where an initial alignment can be obtained through the back-translation corpora generated by an unsupervised SMT system [Artetxe *et al.*, 2019], and (4) language model pre-training [Lample *et al.*, 2018b; Song *et al.*, 2019; Conneau and Lample, 2019], which is discussed in detail in Section 2.4.

**Translation improvement.** The translation quality need to be further improved based on the initial alignment, where iterative back translation is commonly used [Lample *et al.*, 2018a; Lample *et al.*, 2018b; Song *et al.*, 2019]. Some works study on improving the iterative back translation in unsupervised NMT. Sun *et al.* [2019] propose to add a term in the training objective to avoid forgetting the alignment from bilingual word embedding during the interactively training. Moreover, unsupervised SMT can also be utilized to boost the iterative back translation. One approach is to first construct pseudo parallel data by leveraging both the unsupervised SMT and NMT systems for back translation and then train the NMT models with the pseudo parallel data [Lample *et al.*, 2018b; Marie *et al.*, 2019]. In addition, SMT can also act as a posterior regularization to denoise the pseudo parallel data generated by NMT systems [Ren *et al.*, 2019].

### 2.4 Language Model Pre-training

Leveraging monolingual data to pre-train language models is effective for many language understanding and generation tasks [Devlin *et al.*, 2018]. Since NMT requires the capability of both language understanding (e.g., NMT encoder) and generation (e.g., NMT decoder), pre-training language model can be very helpful for NMT, especially low-resource NMT. Previous works on language model pre-training for NMT can be divided into two categories depending on the encoder and decoder in NMT are pre-trained separately or jointly. We then review the works according to the two categories.

**Separate pre-training.** Some works pre-train the encoder or/and the decoder separately. XLM [Conneau and Lample, 2019] initialize the encoder and decoder with separate language models training by a combination of masked language modeling (MLM) [Devlin *et al.*, 2018], where some tokens in the text are masked and the model learns to predict the masked tokens, and translation language modeling (TLM), which extends MLM by concatenating parallel sentence pairs as the input sentences. Rothe *et al.* [2020] investigate to initialize the encoder and decoder with variant models, including BERT [Devlin *et al.*, 2018], GPT-2 [Radford *et al.*, 2019], RoBERTa [Liu *et al.*, 2019] and random initialization. It is observed the best performance on English-Germany by a model with BERT-initialized encoder and randomly initialized decoder, or a model with shared encoder and decoder initialized with BERT. Zhu *et al.* [2020] fuse the representations extracted by BERT to the encoder and decoder via attention mechanisms. A drawback of separately pre-training encoder and decoder is that it cannot well train the encoder-decoder-attention, which is very important in NMT to connect the source and target representations for translation. Therefore, some works propose to jointly pre-train the encoder, decoder and attention for better translation accuracy.

**Joint pre-training.** In order to simultaneously learn to understand the input sentences and improve the language generation capability, as well as jointly pre-train each component in NMT models (encoder, decoder and encoder-decoder-attention), MASS [Song *et al.*, 2019] proposes masked sequence to sequence learning that randomly masks a fragment (several consecutive tokens) in the input sentence of the encoder, and predicts the masked fragment in the decoder. Later, BART [Lewis *et al.*, 2020] proposes to add noises and randomly mask some tokens in the input sentences in the encoder, and learn to reconstruct the original text in the decoder. T5 [Raffel *et al.*, 2020] randomly masks some tokens and replace the consecutive tokens with a single sentinel token.

### 2.5 Exploiting Comparable Corpus

Monolingual data of different languages that refer to the same entity (e.g., English and Chinese Wikipedia pages that describe the same object) can be regarded as comparable corpus, which is easier to be obtained compared to parallel data and contains implicit parallel information for NMT systems. The challenge is how to mine the parallel sentences from the comparable corpus and some approaches are proposed to solve this problem. LASER [Artetxe and Schwenk, 2019] is a toolkit based on cross-lingual sentence embeddings, which is a good choice to mine parallel data [Schwenk *et al.*, 2019a]. Wu *et al.* [2019a] propose to first extract potential aligned target sentences given a source sentence, and then make the target sentences better aligned with the source sentence by revising them via an editing mechanism. A self-supervised learning method is proposed in [Ruiter *et al.*, 2019], where finding semantically aligned sentences is considered as an auxiliary task for translation. Besides mining parallel sentence pairs, Wu *et al.* [2019b] take advantage of the aligned topic distribution for weakly paired documents, which is suitable for documents related to the same event or entity but not implicitly aligned in sentences.

### 2.6 Enhancing With Bilingual Dictionary

The bilingual dictionary can be collected either by human annotation or word embedding based alignment [Zhang *et al.*, 2017], which is much easier to obtain than the bilingual sentences. Since the bilingual dictionary contains only word-level information, it is usually used with monolingual data to improve the translation. Existing works utilizing the bilingual dictionary can be categorized into three ways. First, the bilingual dictionary is used to improve the rare words translation. Zhang and Zong [2016a] build pseudo parallel sentences by translating source-side monolingual sentences (that contain rare words) to target language via SMT (that is built based on the bilingual dictionary). Fadaee *et al.* [2017] augment the parallel data by replacing some words in parallel sentences with rare words. Second, bilingual dictionary can also be used to perform word-by-word translation on monolingual data, and accordingly help to improve the low-resource NMT. Pourdamghani *et al.* [2019] propose a two-step approach, which first translates the source monolingual sentences to translationese sentences word-by-word, and then trains a translationese-to-target model. Zhou *et al.* [2019] augment the parallel training data by first re-ordering monolingual sentences in the target language to match the source language and then obtaining pseudo source sentences via word-by-word translation. Third, a recent study [Duan *et al.*, 2020] propose to close the gap of the embedding spaces between the source and target languages by establishing anchoring points based on dictionary.

### 2.7 Summary and Discussions

Back/forward translation and the joint training on both translation directions utilize monolingual data to improve translation models. Unsupervised NMT uses only monolingual data to get an initial alignment and improve the translation via iterative back translation. Language model pre-training initializes the NMT models with language understanding and generation capability using only monolingual data. Comparable corpora are strong supplements to parallel corpus, from which parallel sentences can be extracted based on language models or translation models. Bilingual dictionary contains word-level parallel information, which is helpful on the alignment between two languages. The above techniques can be combined with each other to gain more in low-resource NMT. For example, back/forward translation and the joint training methods on both translation directions can be applied to any existing translation models, and thus can be easily combined with other techniques in low-resource NMT. Moreover, the pre-trained model can either be fine-tuned to translation task via parallel data that may be extracted from comparable corpora, or used as an initial model for unsupervised NMT.

# 3 Exploiting Data From Auxiliary Languages

Human languages share similarities with each other in several aspects: (1) languages in the same/similar language family or typology may share similar writing script, word vocabulary, word order and grammar, (2) languages can influence each other, and a foreign word from another language can be incorporated into a language as it is (referred as loanword). Accordingly, corpora of related languages can be exploited to assist the translation between a low-resource language pair [Dabre *et al.*, 2020]. The methods to leverage multilingual data into low-resource NMT can be categorized into several types: (1) multilingual training, where the low-resource language pair is jointly trained with other language pairs in one model [Johnson *et al.*, 2017], (2) transfer learning [Zoph *et al.*, 2016], where a parent NMT model usually containing rich-resource language pairs is first trained and then fine-tuned on low-resource language pair, and (3) pivot translation, where one or more pivot languages are selected as a bridge between the source and target languages and in this way the source-pivot and pivot-target data can be exploited to help the source-target translation. In the following subsections, we introduce the works in each category, respectively.

## 3.1 Multilingual Training

Multilingual training enjoys three main advantages. First, training multiple language pairs in a single model through parameter sharing can significantly reduce the cost of model training and maintenance compared with training multiple separate models, and can collectively learn the knowledge from multiple languages to help low-resource languages. Second, low-resource language pairs benefit from related rich-resource languages pairs through joint training. Moreover, multilingual NMT offers the possibility to translate on language pairs that are unseen during training, which is called zero-shot translation. In the following paragraphs, we summarize the works on multilingual training from three perspectives (i.e., parameter sharing, designs for low-resource languages and zero-shot translation).

**Parameter sharing.** There are different ways to share model parameters in multilingual training. First, all the encoder, decoder and attention components are independent among different languages [Dong *et al.*, 2015; Zoph and Knight, 2016]. Second, fully shared encoder, decoder and attention components are considered across languages, where a language-specific token is added in the source sentence to specify the target language [Artetxe and Schwenk, 2019; Johnson *et al.*, 2017; Tan *et al.*, 2019c]. Third, in order to simultaneously exploit the characteristic and commonality of different languages, as well as keeping the model compact, some works consider to partially share the model parameters. Blackwood *et al.* [2018] propose to use a specific attention mechanism in the decoder for each target language and share all the remaining model parameters, which is shown to improve the word alignments. Platanios *et al.* [2018] introduces a contextual parameter generator for each language pair, which generates the parameters of the encoder and decoder based on the source and target language embeddings. Wang *et al.* [2019b] improves the translation quality by using language-sensitive embeddings and attentions, as well as incorporating language-sensitive discriminators in the decoder. Zhang *et al.* [2020a] introduce a linear transformation between the shared encoder and decoder for each target language, which requires only one more weight matrix for an additional target language.

**Designs for low-resource languages.** To better exploit the knowledge from multiple languages to help low-resource languages, a lot of works design to improve the multilingual training from different aspects:

- *Auxiliary language selection.* How to effectively select and utilize auxiliary languages is critical to improve the performance of low-resource language pairs in multilingual NMT. Most works consider to select rich-resource languages in the same language family as auxiliary languages, and achieve significant improvement [Neubig and Hu, 2018]. Tan *et al.* [2019a] propose to cluster the languages based on language embedding, which shows better performance than clustering by language family. Wang and Neubig [2019] focus on translating low-resource languages to English with the help of a target conditioned sampling algorithm, where a target sentence is sampled and the source sentences from all the corresponding parallel sentences in multiple languages are chosen based on language-level and sentence-level similarity.
- *Training sample balance.* Considering the limited model capacity and the various training data sizes among different languages, the model may have a bias to rich-resource languages. Accordingly, balancing the data sizes is important for low-resource languages in multilingual NMT. Temperature based sampling is one promising approach, where the temperature term needs to be manually chosen [Arivazhagan *et al.*, 2019a]. Wang *et al.* [2020] propose a method to automatically weight the training data sizes.
- *Word reordering in auxiliary language.* Pre-ordering the words in auxiliary language sentences to align with the desired low-resource language also brings benefits to low-resource NMT [Murthy V *et al.*, 2019].
- *Monolingual data from auxiliary languages* can also be used to improve the low-resource languages by introducing back translation [Sennrich *et al.*, 2016], cross-lingual pre-training [Liu *et al.*, 2020] and meta-learning [Gu *et al.*, 2018] in multilingual model. Furthermore, multilingual NMT can also be trained with monolingual data only by extending the unsupervised NMT [Sen *et al.*, 2019], or aligning the translations to the same language via different

paths in a multilingual model [Xu *et al.*, 2019].

**Zero-shot translation.** Multilingual training brings the possibility of zero-shot translation. For example, a multilingual NMT model trained on $X \leftrightarrow$ English and $Y \leftrightarrow$ English parallel data is possible to translate between $X$ and $Y$ even if it has never seen the parallel data between $X$ and $Y$. Firat *et al.* [2016] achieve zero-resource translation by first training a multilingual model, then fine-tuning on the pseudo-parallel corpus for the target language pair generated via back translation. Based on a fully shared multilingual NMT model, zero-shot translation shows reasonable quality without any additional steps [Johnson *et al.*, 2017]. Some designs can further improve the zero-shot translation in a fully shared multilingual NMT model: (1) introducing an attentional neural interlingua component between the encoder and decoder [Lu *et al.*, 2018], (2) introducing additional terms to the training objective [Arivazhagan *et al.*, 2019b], and (3) correcting the off-target zero-shot translation issue via a random online back translation algorithm [Zhang *et al.*, 2020a]. There are two important observations: (1) incorporating more languages may provide benefits on zero-shot translation [Aharoni *et al.*, 2019], and (2) the quality of zero-shot between similar languages is quite good [Arivazhagan *et al.*, 2019a].

## 3.2 Transfer Learning

A typical method of transfer learning for low-resource NMT is to first train an NMT model on some auxiliary (usually rich-resource) language pairs, which is called parent model, and then fine-tune all or some of the model parameters on a low-resource language pair, which is called child model [Zoph *et al.*, 2016]. There are three main design aspects in transfer learning: (1) how to select the auxiliary language, (2) how to design the joint vocabulary between the auxiliary and low-resource languages, and (3) how to fine-tune the model for low-resource languages.

**Auxiliary language selection.** A common approach is to select rich-resource languages as auxiliary languages [Zoph *et al.*, 2016], where the languages share the same/similar language family or typology with the given low-resource language performs better [Nguyen and Chiang, 2017]. LANGRANK is a framework to automatically detect the optimal auxiliary language based on typological and corpus statistical information [Lin *et al.*, 2019].

**Joint vocabulary design.** A shared vocabulary including learned sub-words of the auxiliary and the desired low-resource language pairs is commonly used [Nguyen and Chiang, 2017]. However, the shared vocabulary is not suitable for transferring a pre-trained parent model to languages with unseen scripts in the vocabulary. To address this problem, Kim *et al.* [2019a] propose to learn a cross-lingual linear mapping between the embeddings of the unseen language and the bilingual parent model.

**Fine-tuning.** One simple method of fine-tuning is to use a parent model on one rich-resource language to initialize the child model and then fine-tune all the parameters on the low-resource language pair [Zoph *et al.*, 2016]. Some parameters can be fixed while fine-tuning, where Bapna *et al.* [2019] fix the parameters of the parent model and add light-weight residual adapters when fine-tuning. Moreover, besides using a bilingual parent model, a multilingual parent model can also be used, which enjoys two main advantages. First, a low-resource language can benefit from multiple auxiliary languages. Second, considering the limited model capacity of a multilingual model, fine-tuning may force the model to focus on the desired low-resource language, and thus improve the performance. Neubig and Hu [2018] compare different settings when fine-tuning a low-resource NMT model from a multilingual model on many-to-English direction, and come up with the conclusions: (1) Warm start, where the parent model is trained with both auxiliary languages and low-resource language, is better than cold start, where the parent model is trained only on auxiliary languages; (2) Fine-tuning from a universal model containing dozens of languages outperforms fine-tuning from a model with one similar auxiliary language; (3) Fine-tuning with the data of the low-resource language and one similar rich-resource language outperforms fine-tuning with only low-resource language data. In addition, Tan *et al.* [2019b] suggest warm start for many-to-one setting and cold start for one-to-many setting.

## 3.3 Pivot Translation

In pivot-based approaches, a pivot language, which is usually a rich-resource language, is selected as a bridge. Then, the source-pivot and pivot-target corpora and model can be exploited to build the source-target translation.

One approach is to train the source-pivot and pivot-target models and directly combine them as a source-pivot-target model [Cheng *et al.*, 2017]. Another widely used method is to train the source-target model by pseudo-parallel data, which is generated with the help of the pivot language. Zheng *et al.* [2017] translate the pivot language in a pivot-target parallel corpus to source language by a pivot-source NMT model, while Chen *et al.* [2017] build the pseudo-parallel corpora by the source-pivot corpus and pivot-target model. Besides the parallel corpus, the monolingual data of source and target languages can also be used to generate pseudo-parallel corpora [Karakanta *et al.*, 2018; He *et al.*, 2019]. Moreover, leveraging the parameters of source-pivot and pivot-target models is also one way to utilize the pivot language. Kim *et al.* [2019b] transfer the encoder of source-pivot model and the decoder of the pivot-target model to the source-target model [Kim *et al.*, 2019b]. Ji *et al.* [2020] pre-train a universal encoder for source and pivot languages based on cross-lingual pre-training [Conneau and Lample, 2019], and then train on pivot-target parallel data with part of the encoder frozen. Pivot languages selection is critical in pivot-translation, which greatly influences the translation quality. In most cases, one pivot language is selected based on prior knowledge. There also exists a learning to route (LTR) method to automatically select one or several pivot languages to translate via multiple hops [Leng *et al.*, 2019].

## 3.4 Summary and Discussions

Both multilingual training and transfer learning are good ways to learn from auxiliary languages. In multilingual training, a low-resource language is trained with auxiliary lan-

| Dateset | Type | #Language | Size |
|---|---|---|---|
| Wikipedia | mo | 300+ | $\sim$ 55M documents |
| CommonCrawl | mo | 150+ | Billions of URLs |
| CC-100 | mo | 100+ | $\sim$ 0.5B sents/lang |
| JW300 | bi | 300+ | $\sim$ 0.1M sents/pair |
| CCAligned | bi | 137 | $\sim$ 0.3M sents/pair |
| CCMatrix | bi | 80 | $\sim$ 1+M sents/pair |
| WikiMatrix | bi | 85 | $\sim$ 0.1M sents/pair |

Table 1: List of datasets, where mo and bi stand for monolingual and bilingual data, sents/lang and sents/pair stand for the number of sentences in a language and language pair, respectively.

guages from scratch, while in transfer learning, an existing translation model is fine-tuned on a low-resource language. Multilingual training and transfer learning can be combined by fine-tuning from a multilingual model. Pivot translation can be used when the translation path from a source language to a target language can be linked with one or several pivot languages, where each language pair on the path has sufficient training data to ensure high-quality translation. In practice, the methods in Section 2 and 3 can be combined to further improve the translation accuracy on low-resource languages. For example, one can first train a multilingual NMT model, and then fine-tune it to a low-resource language with iterative back and forward translation.

## 4 Exploiting Multi-Modal Data

The parallel data in other modality is also useful for NMT, such as image, video, speech, etc. Chen *et al.* [2019] built a pseudo parallel corpus by generating captions of the same image in both source and target languages via pre-trained image captioning models. In addition, the image caption/description and translation tasks can be jointly learned to incorporate image information [Luong *et al.*, 2015]. Moreover, the image data can be utilized by introducing an additional image component (e.g., encoder, decoder or attention) into the NMT model and aligning the two languages with the corresponding image in the latent space [Su *et al.*, 2019; Nakayama and Nishida, 2017]. Currently, the application of using image-text parallel data on NMT is limited, since such image-text data is always hard to collect for low-resource languages. One potential data source to build new image-text dataset is the images and corresponding captions on websites (e.g., Wikipedia and news pages). For the languages with only speech but no text scripts, speech data can be leveraged to develop the translation capability [Zhang *et al.*, 2020b].

## 5 Datasets

Data is critical for low-resource NMT. In this section, we introduce some corpora that are widely used in low-resource NMT, as shwon in Tab. 1. Wikipedia[2] and Common Crawl[2] contain abundant monolingual data, where Wikipedia covers more than 300 languages and Common Crawl contains billions of web pages crawled from the Internet. CC-100 [Conneau *et al.*, 2020; Wenzek *et al.*, 2020] is a monolingual corpus covering 100+ languages processed from Common Crawl. JW300 [Agić and Vulić, 2019], CCAligned [El-Kishky *et al.*, 2020], CCMatrix [Schwenk *et al.*, 2019b] and WikiMatrix [Schwenk *et al.*, 2019a] extract parallel sentences from monolingual data, where JW300 is from the website jw.org, CCAligned and CCMatrix are aligned from Common Crawl, and WikiMatrix is based on Wikipedia. Moreover, OPUS [Tiedemann, 2012] and HuggingFace[2] provide a collection of open source corpora, which makes it convenient to collect data from multiple data sources.

## 6 Conclusion and Future Directions

In this paper, we provided a literature review for low-resource NMT. Different techniques are classified based on the type of auxiliary data: monolingual data from the source/target languages, data from other languages, and multi-modal data. We hope this survey can help readers to better understand the field and choose appropriate techniques for their applications.

Though lots of efforts have been made on low-resource NMT as surveyed, there still remain some open problems:

- In multilingual and transfer learning, how many and which auxiliary languages to use is unclear. LANGRANK [Lin *et al.*, 2019] trains a model to select one auxiliary language. Intuitively, using multiple auxiliary languages may outperform only one, which is worth exploration.
- Training a multilingual model including multiple rich-resource languages is costly. Transferring a multilingual model to an unseen low-resource language is an efficient approach, where the challenge is how to handle the new vocabulary of the unseen language.
- Bilingual dictionary is useful and easy-to-get. Current works focus on taking advantage of bilingual dictionary on the source and target language. It is also possible to use bilingual dictionary between a low-resource language and auxiliary languages in multilingual and transfer training.
- In terms of multi-modality, speech data has potential to boost NMT, but such studies are limited. For example, some languages are close in speech but different in script (e.g., Tajik and Persian).
- Current approaches have made significant improvements for low-resource languages that either have sufficient monolingual data or are related to some rich-resource languages. Unfortunately, some low-resource languages (e.g., Adyghe and Xibe) have very limited monolingual data and are distant from rich-resource languages. How to handle such languages is challenging and worth further studies.

## References

[Agić and Vulić, 2019] Ž. Agić and I. Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *ACL*, pages 1–44, 2019.

[Aharoni *et al.*, 2019] R. Aharoni, M. Johnson, and O. Firat. Massively multilingual neural machine translation. In *NAACL*, 2019.

[2] Wikipedia: https://www.wikipedia.org/; Common Crawl: http://commoncrawl.org/; HugginFace: https://huggingface.co/

[Arivazhagan *et al.*, 2019a] N. Arivazhagan, A. Bapna, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv:1907.05019*, 2019.

[Arivazhagan *et al.*, 2019b] N. Arivazhagan, A. Bapna, et al. The missing ingredient in zero-shot neural machine translation. *arXiv:1903.07091*, 2019.

[Artetxe and Schwenk, 2019] M. Artetxe and Ho. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, pages 597–610, 2019.

[Artetxe *et al.*, 2018] M. Artetxe, G. Labaka, et al. Unsupervised neural machine translation. In *ICLR*, 2018.

[Artetxe *et al.*, 2019] M. Artetxe, G. Labaka, et al. An effective approach to unsupervised machine translation. In *ACL*, 2019.

[Bahdanau *et al.*, 2014] D. Bahdanau, K. H. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. 2014.

[Bapna *et al.*, 2019] A. Bapna, N. Arivazhagan, and O. Firat. Simple, scalable adaptation for neural machine translation. In *EMNLP-IJCNLP*, pages 1538–1548, 2019.

[Blackwood *et al.*, 2018] G. Blackwood, M. Ballesteros, and T. Ward. Multilingual neural machine translation with task-specific attention. In *COLING*, pages 3112–3122, 2018.

[Caswell *et al.*, 2019] I. Caswell, C. Chelba, and D. Grangier. Tagged back-translation. In *WMT*, pages 53–63, 2019.

[Chen *et al.*, 2017] Y. Chen, Y. Liu, et al. A teacher-student framework for zero-resource neural machine translation. In *ACL*, 2017.

[Chen *et al.*, 2019] S. Chen, Q. Jin, and J. Fu. From words to sentences: A progressive learning approach for zero-resource machine translation with visual pivots. In *IJCAI*, 2019.

[Cheng *et al.*, 2017] Y. Cheng, Q. Yang, et al. Joint training for pivot-based neural machine translation. In *IJCAI*, 2017.

[Chu and Wang, 2018] C. Chu and R. Wang. A survey of domain adaptation for neural machine translation. In *COLING*, pages 1304–1319, 2018.

[Conneau and Lample, 2019] A. Conneau and G. Lample. Cross-lingual language model pretraining. *arXiv:1901.07291*, 2019.

[Conneau *et al.*, 2020] A. Conneau, K. Khandelwal, et al. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451, 2020.

[Dabre *et al.*, 2020] R. Dabre, C. Chu, and A. Kunchukuttan. A survey of multilingual neural machine translation. *CSUR*, 2020.

[Devlin *et al.*, 2018] J. Devlin, M. Chang, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[Dong *et al.*, 2015] D. Dong, H. Wu, et al. Multi-task learning for multiple language translation. In *ACL-IJCNLP*, 2015.

[Duan *et al.*, 2020] X. Duan, B. Ji, et al. Bilingual dictionary based neural machine translation without using parallel sentences. In *ACL*, pages 1570–1579, 2020.

[Edunov *et al.*, 2018] S. Edunov, M. Ott, et al. Understanding back-translation at scale. In *EMNLP*, pages 489–500, 2018.

[El-Kishky *et al.*, 2020] A. El-Kishky, V. Chaudhary, et al. A massive collection of cross-lingual web-document pairs. In *EMNLP*, pages 5960–5969, 2020.

[Fadaee *et al.*, 2017] M. Fadaee, A. Bisazza, et al. Data augmentation for low-resource neural machine translation. In *ACL*, 2017.

[Firat *et al.*, 2016] O. Firat, B. Sankaran, et al. Zero-resource translation with multi-lingual neural machine translation. In *EMNLP*, pages 268–277, 2016.

[Gehring *et al.*, 2017] J. Gehring, M. Auli, et al. A convolutional encoder model for neural machine translation. In *ACL*, 2017.

[Gu *et al.*, 2018] J. Gu, Y. Wang, et al. Meta-learning for low-resource neural machine translation. In *EMNLP*, 2018.

[He *et al.*, 2016] D. He, Y. Xia, et al. Dual learning for machine translation. In *NIPS*, pages 820–828, 2016.

[He *et al.*, 2019] T. He, J. Chen, et al. Language graph distillation for low-resource machine translation. *arXiv:1908.06258*, 2019.

[Hoang *et al.*, 2018] V. C. D. Hoang, P. Koehn, et al. Iterative back-translation for neural machine translation. In *WNMT*, 2018.

[Imamura *et al.*, 2018] K. Imamura, A. Fujita, et al. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *ACL*, pages 55–63, 2018.

[Ji *et al.*, 2020] B. Ji, Z. Zhang, et al. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *AAAI*, pages 115–122, 2020.

[Johnson *et al.*, 2017] M. Johnson, M. Schuster, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, pages 339–351, 2017.

[Karakanta *et al.*, 2018] A. Karakanta, J. Dehdari, and J. van Genabith. Neural machine translation for low-resource languages without parallel corpora. *MT*, pages 167–189, 2018.

[Kim *et al.*, 2019a] Y. Kim, Y. Gao, and H. Ney. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. *ACL*, pages 1246–1257, 2019.

[Kim *et al.*, 2019b] Y. Kim, P. Petrov, et al. Pivot-based transfer learning for neural machine translation between non-english languages. In *EMNLP-IJCNLP*, pages 865–875, 2019.

[Lample *et al.*, 2018a] G. Lample, A. Conneau, et al. Unsupervised machine translation using monolingual corpora only. In *ICLR*, 2018.

[Lample *et al.*, 2018b] G. Lample, M. Ott, et al. Phrase-based & neural unsupervised machine translation. In *EMNLP*, pages 5039–5049, 2018.

[Leng *et al.*, 2019] Y. Leng, X. Tan, et al. Unsupervised pivot translation for distant languages. In *ACL*, pages 175–183, 2019.

[Lewis *et al.*, 2020] M. Lewis, Y. Liu, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[Lin *et al.*, 2019] Y. Lin, C. Chen, et al. Choosing transfer languages for cross-lingual learning. *ACL*, 2019.

[Liu *et al.*, 2019] Y. Liu, M. Ott, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.

[Liu *et al.*, 2020] Y. Liu, J. Gu, et al. Multilingual denoising pre-training for neural machine translation. *TACL*, 2020.

[Lu *et al.*, 2018] Y. Lu, P. Keung, et al. A neural interlingua for multilingual machine translation. In *WMT*, pages 84–92, 2018.

[Luong *et al.*, 2015] M. Luong, Q. V Le, et al. Multi-task sequence to sequence learning. *arXiv:1511.06114*, 2015.

[Marie *et al.*, 2019] B. Marie, H. Sun, et al. Nict's unsupervised neural and statistical machine translation systems for the wmt19 news translation task. In *WMT*, pages 294–301, 2019.

[Mikolov *et al.*, 2013] T. Mikolov, I. Sutskever, et al. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 1–9, 2013.

[Murthy V *et al.*, 2019] R. Murthy V, A. Kunchukuttan, and P. Bhattacharyya. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *NAACL*, pages 3868–3873, 2019.

[Nakayama and Nishida, 2017] H. Nakayama and N. Nishida. Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *MT*, pages 49–64, 2017.

[Neubig and Hu, 2018] G. Neubig and J. Hu. Rapid adaptation of neural machine translation to new languages. In *EMNLP*, 2018.

[Nguyen and Chiang, 2017] T. Q Nguyen and D. Chiang. Transfer learning across low-resource, related languages for neural machine translation. In *IJCNLP*, pages 296–301, 2017.

[Nguyen *et al.*, 2019] X.P. Nguyen, S. Joty, et al. Data diversification: A simple strategy for neural machine translation. *arXiv:1911.01986*, 2019.

[Niu *et al.*, 2018] X. Niu, M. Denkowski, and M. Carpuat. Bidirectional neural machine translation with synthetic parallel data. In *NGT*, pages 84–91, 2018.

[Platanios *et al.*, 2018] E. A. Platanios, M. Sachan, et al. Contextual parameter generation for universal neural machine translation. In *EMNLP*, pages 425–435, 2018.

[Pourdamghani *et al.*, 2019] N. Pourdamghani, N. Aldarrab, et al. Translating translationese: A two-step approach to unsupervised machine translation. In *ACL*, pages 3057–3062, 2019.

[Qin, 2020] T. Qin. *Dual Learning*. Springer, 2020.

[Radford *et al.*, 2019] A. Radford, J. Wu, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[Raffel *et al.*, 2020] C. Raffel, N. Shazeer, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, pages 1–67, 2020.

[Ren *et al.*, 2019] S. Ren, Z. Zhang, et al. Unsupervised neural machine translation with smt as posterior regularization. In *AAAI*, pages 241–248, 2019.

[Rothe *et al.*, 2020] S. Rothe, S. Narayan, and A. Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *TACL*, pages 264–280, 2020.

[Ruiter *et al.*, 2019] D. Ruiter, C. Espana-Bonet, and J. van Genabith. Self-supervised neural machine translation. In *ACL*, 2019.

[Schwenk *et al.*, 2019a] H. Schwenk, V. Chaudhary, et al. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv:1907.05791*, 2019.

[Schwenk *et al.*, 2019b] H. Schwenk, G. Wenzek, et al. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv:1911.04944*, 2019.

[Sen *et al.*, 2019] S. Sen, K. K. Gupta, et al. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *ACL*, pages 3083–3089, 2019.

[Sennrich *et al.*, 2016] R. Sennrich, B. Haddow, et al. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96, 2016.

[Song *et al.*, 2019] K. Song, X. Tan, et al. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*, 2019.

[Su *et al.*, 2019] Y. Su, K. Fan, et al. Unsupervised multi-modal neural machine translation. In *CVPR*, pages 10482–10491, 2019.

[Sun *et al.*, 2019] H. Sun, R. Wang, et al. Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *ACL*, 2019.

[Tan *et al.*, 2019a] X. Tan, J. Chen, et al. Multilingual neural machine translation with language clustering. In *EMNLP-IJCNLP*, pages 962–972, 2019.

[Tan *et al.*, 2019b] X. Tan, Y. Leng, et al. A study of multilingual neural machine translation. *arXiv:1912.11625*, 2019.

[Tan *et al.*, 2019c] X. Tan, Y. Ren, et al. Multilingual neural machine translation with knowledge distillation. *ICLR*, 2019.

[Tiedemann, 2012] J. Tiedemann. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218, 2012.

[Vaswani *et al.*, 2017] A. Vaswani, N. Shazeer, et al. Attention is all you need. In *NIPS*, pages 6000–6010, 2017.

[Vincent *et al.*, 2008] P. Vincent, H. Larochelle, et al. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008.

[Wang and Neubig, 2019] X. Wang and G. Neubig. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *ACL*, pages 5823–5828, 2019.

[Wang *et al.*, 2019a] Y. Wang, Y. Xia, et al. Multi-agent dual learning. In *ICLR*, 2019.

[Wang *et al.*, 2019b] Y. Wang, L. Zhou, et al. A compact and language-sensitive multilingual translation method. In *ACL*, pages 1213–1223, 2019.

[Wang *et al.*, 2020] X. Wang, Y. Tsvetkov, et al. Balancing training for multilingual neural machine translation. In *ACL*, 2020.

[Wenzek *et al.*, 2020] G. Wenzek, M. Lachaux, et al. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *COLING*, pages 4003–4012, 2020.

[Wu *et al.*, 2019a] J. Wu, X. Wang, and Y. Wang. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. In *NAACL*, pages 1173–1183, 2019.

[Wu *et al.*, 2019b] L. Wu, J. Zhu, et al. Machine translation with weakly paired documents. In *EMNLP-IJCNLP*, 2019.

[Xu *et al.*, 2019] C. Xu, T. Qin, et al. Polygon-net: A general framework for jointly boosting multiple unsupervised neural machine translation models. In *IJCAI*, pages 5320–5326, 2019.

[Yang *et al.*, 2018] Z. Yang, W. Chen, et al. Unsupervised neural machine translation with weight sharing. In *ACL*, 2018.

[Zhang and Zong, 2016a] J. Zhang and C. Zong. Bridging neural machine translation and bilingual dictionaries. *arXiv:1610.07272*, 2016.

[Zhang and Zong, 2016b] J. Zhang and C. Zong. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*, pages 1535–1545, 2016.

[Zhang *et al.*, 2017] M. Zhang, Y. Liu, et al. Adversarial training for unsupervised bilingual lexicon induction. In *ACL*, 2017.

[Zhang *et al.*, 2020a] B. Zhang, P. Williams, et al. Improving massively multilingual neural machine translation and zero-shot translation. In *ACL*, pages 1628–1639, 2020.

[Zhang *et al.*, 2020b] C. Zhang, X. Tan, et al. Uwspeech: Speech to speech translation for unwritten languages. *arXiv:2006.07926*, 2020.

[Zheng *et al.*, 2017] H. Zheng, Y. Cheng, and Y. Liu. Maximum expected likelihood estimation for zero-resource neural machine translation. In *IJCAI*, pages 4251–4257, 2017.

[Zheng *et al.*, 2020] Z. Zheng, H. Zhou, et al. Mirror-generative neural machine translation. In *ICLR*, 2020.

[Zhou *et al.*, 2019] C. Zhou, X. Ma, et al. Handling syntactic divergence in low-resource machine translation. In *EMNLP-IJCNLP*, pages 1388–1394, 2019.

[Zhu *et al.*, 2020] J. Zhu, Y. Xia, et al. Incorporating bert into neural machine translation. *arXiv:2002.06823*, 2020.

[Zoph and Knight, 2016] B. Zoph and K. Knight. Multi-source neural translation. In *NAACL*, pages 30–34, 2016.

[Zoph *et al.*, 2016] B. Zoph, D. Yuret, et al. Transfer learning for low-resource neural machine translation. In *EMNLP*, 2016.