A Survey of Multilingual Neural Machine Translation

RAJ DABRE, National Institute of Information and Communications Technology (NICT), Kyoto, Japan
CHENHUI CHU, Osaka University, Osaka, Japan
ANOOP KUNCHUKUTTAN, Microsoft, Hyderabad, India

We present a survey on multilingual neural machine translation (MNMT), which has gained a lot of traction in recent years. MNMT has been useful in improving translation quality as a result of translation knowledge transfer (transfer learning). MNMT is more promising and interesting than its statistical machine translation counterpart, because end-to-end modeling and distributed representations open new avenues for research on machine translation. Many approaches have been proposed to exploit multilingual parallel corpora for improving translation quality. However, the lack of a comprehensive survey makes it difficult to determine which approaches are promising and, hence, deserve further exploration. In this article, we present an in-depth survey of existing literature on MNMT. We first categorize various approaches based on their central use-case and then further categorize them based on resource scenarios, underlying modeling principles, core-issues, and challenges. Wherever possible, we address the strengths and weaknesses of several techniques by comparing them with each other. We also discuss the future directions for MNMT. This article is aimed towards both beginners and experts in NMT. We hope this article will serve as a starting point as well as a source of new ideas for researchers and engineers interested in MNMT.

CCS Concepts: • **Computing methodologies** → **Machine translation**;

Additional Key Words and Phrases: Neural machine translation, survey, multilingualism, low-resource, zero-shot, multi-source

ACM Reference format:
Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.* 53, 5, Article 99 (September 2020), 38 pages.
https://doi.org/10.1145/3406095

1 INTRODUCTION

Neural machine translation (NMT) [8, 24, 140] has become the dominant paradigm for MT in academic research as well as commercial use [161]. NMT has shown state-of-the-art performance for many language pairs [14, 15]. Its success can be mainly attributed to the use of distributed

All authors contributed equally to this research.
A part of this work was conducted under the program "Research and Development of Enhanced Multilingual and Multipurpose Speech Translation System" of the Ministry of Internal Affairs and Communications (MIC), Japan. This work was also partially supported by Grant-in-Aid for Young Scientists #19K20343, JSPS.
Authors' addresses: R. Dabre, National Institute of Information and Communications Technology (NICT), 3-5, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan; email: prajdabre@gmail.com; C. Chu, Institute for Datability Science, Osaka University, C503, Technoalliance Building, Osaka University, 2-8 Yamadaoka, Suita, Osaka, 565-0871, Japan; email: chu@ids.osaka-u.ac.jp; A. Kunchukuttan, Microsoft (R&D) India Pvt. Ltd., Microsoft Campus, Gachibowli, Hyderabad, Telangana, India, PIN: 500-032; email: anoop.kunchukuttan@microsoft.com.

0360-0300/2020/09-ART99
https://doi.org/10.1145/3406095

representations of language, enabling end-to-end training of an MT system. Unlike classical statistical machine translation (SMT) systems [79], separate lossy components such as word aligners, translation rule extractors, and other feature extractors are not required. The dominant NMT approach is the *Embed - Encode - Attend - Decode* paradigm. Recurrent neural network (RNN) [8], convolutional neural network (CNN) [52], and self-attention/feed-forward network (SA/FFN) [148] architectures are most commonly used approaches based on this paradigm. For a more detailed exposition of NMT, we refer readers to some prominent tutorials [78, 104].

While initial research on NMT started with building translation systems between two languages, researchers discovered that the NMT framework can naturally incorporate multiple languages. Hence, there has been a massive increase in work on MT systems that involve more than two languages [21–23, 42, 48, 70, 105, 168]. We refer to NMT systems handling translation between more than one language pair as *multilingual NMT* (MNMT) systems. The ultimate goal of MNMT research is to develop one model for translation between as many languages as possible by effective use of available linguistic resources.

MNMT systems are desirable, because training models with data from many language pairs might help a resource-poor language acquire extra knowledge from the other languages [134, 168]. Moreover, MNMT systems tend to generalize[1] better due to exposure to diverse languages, leading to improved translation quality compared to bilingual NMT systems. This particular phenomenon is known as translation knowledge transfer [112]. For the rest of the article, we will use the terms "knowledge transfer," "transfer learning," and "translation knowledge transfer" interchangeably. Knowledge transfer has been strongly observed for translation between low-resource languages, which have scarce parallel corpora or other linguistic resources but have benefited from data in other languages [170]. Knowledge transfer also has been observed between high-resource languages and multi-source NMT, where MNMT systems outperform bilingual NMT systems [49, 70, 168]. Multilingual training is also known to be a source of regularization during training, which further improves generalization. In addition, MNMT systems have the potential to be relatively compact,[2] because a single model handles translations between multiple language pairs [70]. This can reduce the deployment footprint, which is crucial for constrained environments such as mobile phones or IoT devices. It can also simplify the large-scale deployment of MT systems.

There are multiple scenarios where MNMT has been put to use based on available resources and use-cases. The following are the major scenarios where MNMT has been explored in the literature (see Figure 1 for an overview):

Multiway Translation. The goal is constructing a single NMT system for one-to-many [42], many-to-one [93], or many-to-many [48] translation using parallel corpora for more than one language pair. In this scenario, we make a very generic assumption that parallel corpora for a number of languages are available. Here, the ultimate objective is to incorporate a number of languages into a single model.

Low-resource Translation. Little to no parallel corpora exist for most language pairs in the world. Multiple studies have explored using *assisting* languages to improve translation between low-resource language pairs. These multilingual NMT approaches for low-resource MT address two broad scenarios: (a) a high-resource language pair (e.g., Spanish-English) is available to assist a low-resource language pair (e.g., Catalan-English). Transfer learning is typically used in this scenario [170]; (b) no direct parallel corpus for the low-resource pair, but languages share a parallel corpus with one or more *pivot* language(s).

[1]Using additional languages can help in word sense disambiguation, which can help improve translation quality.
[2]Relatively modest increase in model parameters despite an increase in the number of languages as a result of parameter sharing.

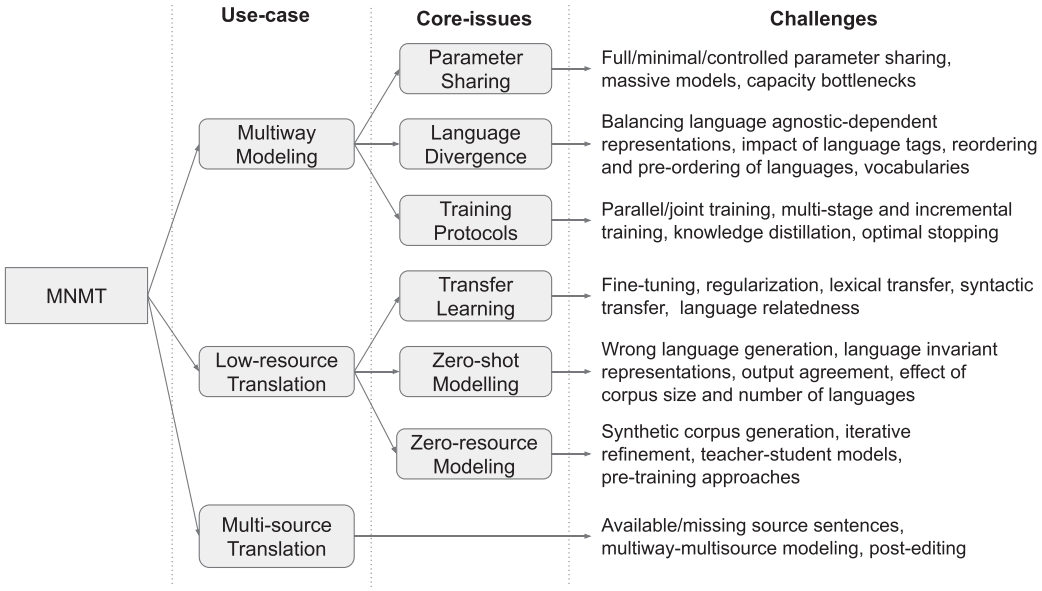

Fig. 1. MNMT research categorized according to use-cases, core-issues, and the challenges involved. Note that the focus is on use-cases and an approach for one use-case can be naturally adapted to another use-case and thus can face similar core issues.

Multi-source Translation. Documents that have been translated into more than one language might, in the future, be required to be translated into another language. In this scenario, existing multilingual complementary content on the source side can be exploited for multi-source translation [168]. Multilingual complementary content can help in better disambiguation of content to be translated, leading to an improvement in translation quality.

We believe that the biggest benefit of doing MNMT research by exploring these scenarios is that we might gain insights and an answer to an important question in natural language processing (NLP):

Q. *How can we leverage multilingual data effectively to learn distributions across multiple languages to improve MT (NLP) performance across all languages?*

This question can be decomposed into a number of smaller questions, and in this article, we try to answer two of them as follows:

Q1. *Is it possible to have a one-model-for-all-languages solution to MT (NLP) applications?*
Q2. *Can shared multilingual distributed representations help MT (NLP) for low-resource languages?*

Given these benefits, scenarios, and the tremendous increase in the work on MNMT in recent years, we write this survey article on MNMT to systematically organize the work in this area. To the best of our knowledge, no such comprehensive survey on MNMT exists. Our goal is to shed light on various MNMT scenarios, fundamental questions in MNMT, basic principles, architectures, and datasets for MNMT systems. The remainder of this article is structured as follows: We present a systematic categorization of different approaches to MNMT in each of the above-mentioned scenarios to help understand the array of design choices available while building MNMT systems (Sections 3, 4, 5, and 6). The top-level organization of the survey is use-case-scenario-based: multiway NMT (Section 3), low-resource NMT (Sections 4 and 5), and multi-source NMT (Section 6).

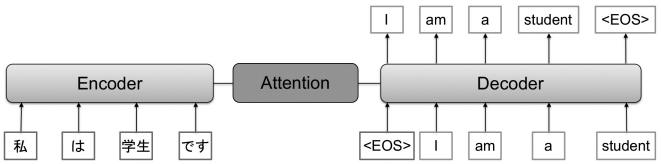

Fig. 2. A standard NMT model based on the encode-attend-decode modeling approach.

Although zero-shot/zero-resource is a special case of low-resource NMT, we dedicate a separate section (5) given its growing importance and interest. For each scenario, we address the challenges, considerations, and multilingual NMT-based solutions. Note that a model or technique proposed for one scenario may be used in another scenario. We have mentioned proposed solutions in the context of the scenario in which they have been discussed in the literature. Their applicability to other scenarios may be subject to further investigation. We put the work in MNMT into a historical perspective with respect to multilingual MT in older MT paradigms (Section 7). We also describe most commonly used multilingual datasets and the shared tasks that focus on multilingual NMT (Section 8). In addition, we compare MNMT with domain adaptation for NMT, which tackles the problem of improving low-resource in-domain translation (Section 9). Finally, we share our opinions on future research directions in MNMT (Section 10) and conclude this article (Section 11).

2 NEURAL MACHINE TRANSLATION

Given a parallel corpus $\mathbf{C}$ consisting of a set of parallel sentence pairs $(\mathbf{x}, \mathbf{y})$, the training objective for NMT is maximize the log-likelihood $\mathcal{L}$ w.r.t θ:

$$\mathcal{L}_\theta = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{C}} \log p(\mathbf{y}|\mathbf{x};\theta), \tag{1}$$

where $\mathbf{x} = \{x_1, \ldots, x_n\}$ is an input sentence, $\mathbf{y} = \{y_1, \ldots, y_m\}$ is its translation, and θ is a set of parameters to be learned. The probability of a target sentence given the source sentence is:

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{j=1}^{m} p(y_j|y_{<j}, \mathbf{x};\theta),$$

where m is the number of words in $\mathbf{y}$, y_j is the current generated word, and $y_{<j}$ are the previously generated words. At inference time, beam search is typically used to find the translation that maximizes the above probability.

The most commonly used NMT approach is the *Embed - Encode - Attend - Decode* paradigm. Figure 2 shows an overview of this paradigm. The encoder first converts words in the source sentence into word embeddings. These word embeddings are then processed by neural layers and converted to representations that capture contextual information about these words. We call these contextual representations as the *encoder representations*. The decoder uses an attention mechanism, the encoder representations, and previously generated words to generate what we call the decoder representations (states), which in turn are used to generate the next target word. The encoder and decoder can be RNN [8], CNN [52], or self-attention and feed-forward [148] layers. Among these, the self-attention layers are the most widely used. It is a common practice to stack multiple layers, which leads to an improvement in translation quality. The attention mechanism is calculated cross the decoder and encoder as:

$$e_{ji} = a(\mathbf{s_{j-1}}, \mathbf{h_i}), \tag{2}$$

$$a_{ji} = \frac{\exp(e_{ji})}{\sum_{k=1}^{m} \exp(e_{ki})},$$

where e_{ji} is an alignment score, a is an alignment model that scores the match level of the inputs around position i and the output at position j, $\mathbf{s_{j-1}}$ is the decoder hidden state of the previous generated word, $\mathbf{h_i}$ is the encoder hidden state at position i. The calculated attention vector is then used to weight the encoder hidden states to obtain a context vector as:

$$\mathbf{c_j} = \sum_{i=1}^{n} a_{ji}\mathbf{h_i}.$$

This context vector, is fed to the decoder along with the previously generated word and its hidden state to produce a representation for generating the current word. A decoder hidden state for the current word $\mathbf{s_j}$ is computed by:

$$\mathbf{s_j} = g(\mathbf{s_{j-1}}, \mathbf{y_{j-1}}, \mathbf{c_j}),$$

where g is an activation decoder function, $\mathbf{s_{j-1}}$ is the previous decoder hidden state, $\mathbf{y_{j-1}}$ is the embedding of the previous word. The current decoder hidden state $\mathbf{s_j}$, the previous word embedding, and the context vector are fed to a feedforward layer f and a softmax layer to compute a score for generating a target word as output:

$$P(y_j|y_{<j}, \mathbf{x}) = \text{softmax}(f(\mathbf{s_j}, \mathbf{y_{j-1}}, \mathbf{c_j})).$$

Training NMT models. The parallel corpus used to train the NMT model is first subjected to pre-processing, where it is sufficiently cleaned to remove noisy training examples. A vocabulary of the N most frequent words is then created and the remaining words are treated as unknown words mapped to a single token designated by "UNK." To overcome the problem of unknown words, the most common practice involves subword tokenization using methods such as byte-pair encoding (BPE) [131], word-piece model (WPM) [127], or sentence-piece model (SPM) [82]. This enables the use of an open vocabulary. To train an NMT model, we typically minimize the cross-entropy[3] (loss) between the predicted target words and the actual target words in the reference. This loss minimization is an optimization problem and gradient descent methods such as SGD, ADAM, ADAGRAD, Adafactor, and so on, can be used [122]. ADAM is widely used in MT due to its ability to quickly train models but suffers from the inability to sufficiently converge. However, SGD is known to converge better but requires long training times. Designing a learning schedule that combines several optimizers can help train a model with high performance [161]. However, following the popular work on the Transformer model [148], most recent research focuses on using a single optimizer with a learning rate schedule where the learning rate is altered depending on the training iteration. Training is either done for a large number of iterations or till the model converges sufficiently. Typically, a model is said to converge when its evaluation on a development set does not change by a significant amount over several iterations. We refer readers to previous works on NMT regarding this topic. Another consideration during training is the tuning of hyper-parameters such as learning rate, hidden dimension size, number of layers, and so on. Researchers often train a wide variety of models, an approach known as hyperparameter search, and choose a model with the best performance. People who are familiar with deep learning might notice that this basic model can be naturally extended to involve multiple language pairs.

Decoding NMT models. Once an NMT model has been trained, it can be used to translate a sentence into another language. This process is also known as translation, decoding, or inference.

[3]This also implies maximizing the likelihood of the predicted sequence of words in the target language.

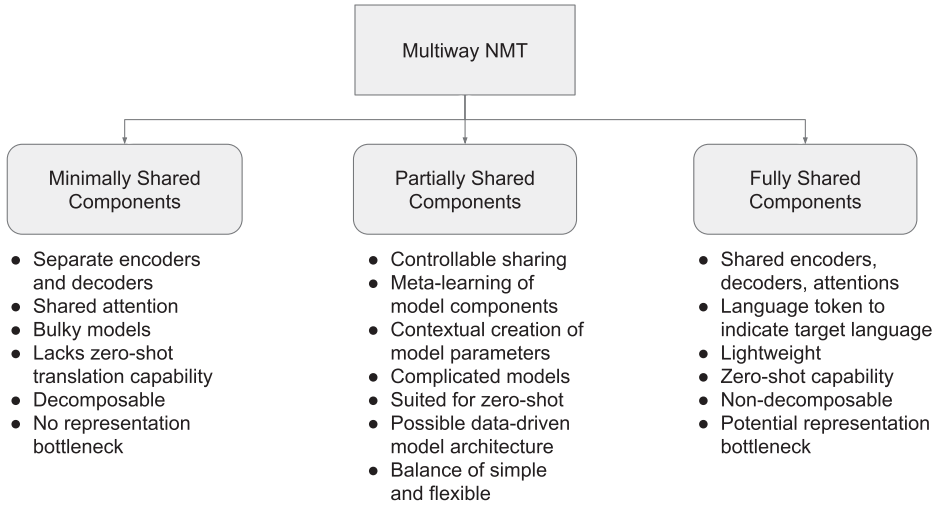

Fig. 3. An overview of multiway NMT from the perspective of the level of sharing and the features and limitations each sharing approach. All types of MNMT models have to deal with complex training issues ranging from batching to language grouping to knowledge distillation. Additionally, it is important to address language divergence and finding the right balance of language-specific and language-agnostic representations.

The simplest decoding algorithm is beam search decoding. Typically, the development set source sentences are decoded using combinations of beam size and a length penalty (also referred to as alpha) and the combination that gives the best automatic evaluation metric score is chosen to decode the test set.

3 MULTIWAY NMT

The primary goal of MNMT is a model that can support translation between more than one language pair. Refer to Figure 3 for an overview of the multiway NMT paradigm. We use the term *multiway* NMT models to denote such models. Formally, a single model can support translation for l language pairs $(src_l, tgt_l) \in \mathbf{L} \quad (l = 1 \text{ to } L)$, where $\mathbf{L} \subset S \times T$, and S, T are sets of X source and Y target languages, respectively. S and T need not be mutually exclusive. Parallel corpora are available for **all** of these l language pairs as $C(src_l)$ and $C(tgt_l)$. Note that our objective in this specific scenario is to train a translation system between all language pairs. Refer to Figure 4 for the two prototypical MNMT approaches with minimal and complete sharing of components. Most existing works are variations of these models. Particularly, one-to-many [42], many-to-one [93] and many-to-many [48] NMT models are specific instances of this general framework. The training objective for multiway NMT is maximization of the log-likelihood of all training data jointly for all language pairs (different weights may be assigned to the likelihoods of different pairs:):

$$\mathcal{L}_\theta = \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}^{C(src_l), C(tgt_l)}(\theta),$$

which can be calculated in the same way as Equation (1).

Multiway NMT systems are of great interest, since it is believed that transfer learning between languages can take place that will help improve the overall translation quality for many translation directions [170] and at the same time enable translations between language pairs with no data

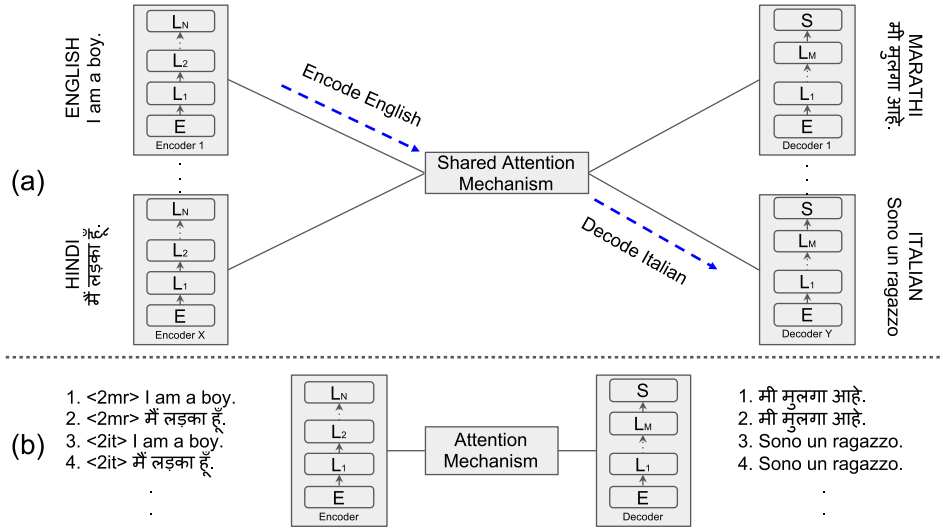

Fig. 4. Part (a) of the figure shows an X sources (X encoders) and Y targets (Y decoders) MNMT model. The encoders contain N layers (L^1 to L^N) and decoders contain M layers ((L^1 to L^M)) that can be stacked. Each layer can be recurrent, convolutional, or feed-forward. The decoders contain a softmax layer (S) to predict target language words. A single attention mechanism is shared among all encoders and decoders. In the figure, the blue arrows indicate the flow of processing to translate an English sentence into its Italian translation. Given the vast number of components, they can be initialized by pre-trained models such as BERT to perform transfer learning. Post-training, this model can be decomposed into individual bilingual models. Part (b) of the figure shows a fully shared MNMT model for all language pairs. The "language tag" trick where a token like "$< 2xx >$" is prefixed to each source sentence to indicate the desired target language. All existing MNMT models are adaptations of these two fundamental architectures. Depending on the language pairs involved, the size of the data, the languages of focus, and deployment conditions, one should consider choosing the components that should be shared.

[70]. Analyzing multiway NMT systems could also provide an understanding of the relationship between languages from a statistical and linguistic point of view [36, 105].

Multiway translation systems follow the standard embed-encode-attend-decode paradigm that is common for existing NMT systems. However, the architecture is adapted to support multiple languages. This involves addressing issues related to vocabularies and associated embeddings, network layers (RNN/CNN/Feed-Forward), parameter sharing, training protocols and language divergence. We address each issue in this section.

3.1 Parameter Sharing

There are a wide range of architectural choices in the design of MNMT models. The choices are primarily defined by the degree of parameter sharing among various supported languages.

Minimal Parameter Sharing. Firat et al. [48] proposed a model comprised of separate embeddings, encoders and decoders for each language that all shared a single attention mechanism. Additionally there are two shared components: a layer for all encoders for initializing the initial decoder state by using the final encoder state and an affine layer for all decoders to project the final decoder state before computing softmax. However, the focus is on the shared attention layer as it has to bear most of the burden of connecting all source and target languages. Figure 4(a) depicts a simplified view of this model. Different from the attention score calculated for a single language

pair in Equation (2), the attention score in Firat et al. [48] is calculated from multiple encoders and decoders as:

$$e_{ji}^{vu} = a\left(s_{j-1}^{v}, h_i^u\right),$$

where u is the u-th encoder and v is the v-th decoder. By sharing attention across languages, it was hypothesized that transfer learning could take place and such a model was able to outperform bilingual models, especially when the target language was English. However, this model has a large number of parameters, usually around 270M or more. Furthermore, the number of parameters only grows linearly with the number of languages, while it grows quadratically for bilingual systems spanning all the language pairs in the multiway system. Another problem is that the shared attention mechanism has to bear the burden of connecting different language pairs and this can introduce a representational bottleneck where a model cannot learn the necessary representations for the best translation quality. However, we emphasize that minimally shared models are quite flexible because it is possible to treat individual languages with special processing mechanisms especially suited to maximize performance for that language. However after Johnson et al. [70] showed that such models with large number of parameters are not required, research efforts started focusing on a middle-ground where the amount of sharing is controlled.

Complete Parameter Sharing. Johnson et al. [70] proposed a highly compact model where all languages share the same embeddings, encoder, decoder, and attention mechanism. Typically, a common vocabulary across all languages is first generated using a subword-level encoding mechanism such as BPE, WPM, or SPM. Thereafter, all corpora are concatenated[4] and the input sequences are prefixed with a special token (called the *language tag*) to indicate the target language (see Figure 4(b)). This enables the decoder to correctly generate the target language despite all target languages sharing the same decoder parameters. Note that the embedding and softmax layers are shared across all languages and Ha et al. [58] proposed a similar model, but they maintained separate vocabularies for each language. While this might help in faster inference due to smaller softmax layers, the possibility of cognate sharing is lower, especially for linguistically close languages sharing a common script. According to us, Johnson et al. [70] should be particularly useful for related languages, especially those that use the same script, because they have high degree of lexical and syntactic similarity [123]. However, Ha et al. [58] can be used for unrelated languages. In practice, we recommend experimenting with both Ha et al. [58] and Reference [70] and then choosing the appropriate model. We believe that an empirical analysis of these architectures in a variety of language and corpora settings will be helpful. A few techniques through which lexical similarity can be further leveraged are:

- representing all languages in a common script using script conversion [38, 93] or transliteration (Nakov and Ng [103] for multilingual SMT);
- using a common subword-vocabulary across all languages, e.g., character [93] and BPE [106];
- representing words by both character encoding and a latent embedding space shared by all languages [155].

Concerning the choice of non-embedding or softmax layers, Rikters et al. [121] and Lakew et al. [89] have compared RNN, CNN, and the self-attention-based architectures for MNMT. They show that self-attention-based architectures outperform the other architectures in many cases. The most

[4]It is common to oversample smaller corpora so all language pairs are equally represented in the model regardless of the size of the corpora for those pairs.

common hypothesis is that self-attention enables random access to all words in a sentence, which leads to better word and sentence representations.

The complete sharing approach treats the NMT system as a *black box* to train a multilingual system. We would like to emphasize that this type of model has maximum simplicity and has minimal parameter size, usually around 30M (similar to bilingual models), as all languages share the same parameters; and achieves comparable/better results w.r.t. bilingual systems. Work on **massively multilingual NMT** [2, 5, 11] pushes the idea of completely shared models for all language pairs to its limits. Aharoni et al. [2], Arivazhagan et al. [5] trained a single model for 59 and 103 languages and explore a wide range of model configurations focusing on data selection,[5] corpora balancing, vocabulary, deep stacking, training, and decoding approaches. While massively multilingual models have a strong positive impact on low-resource languages, they tend to benefit translation into English a lot more than from English. Furthermore, the gains in translation quality tend to taper off when using more than 50 languages. These works focus on translation performance between language pairs for which there was no training data but we will discuss this separately in Section 5. However, a massively multilingual system also runs into **representation bottlenecks** [2, 135], where not all translation directions show improved performance despite a massive amount of data being fed to a model with a massive number of parameters. Future research should focus on better addressing the bottleneck issues, issues with deeply stacked models, and designing new modular layers that can handle a wide variety and a large number of languages. We strongly recommend readers to read the paper by Arivazhagan et al. [5], which itself is a survey on multilingual NMT on web-scale data. From the perspective of understanding the working of such models, a toolkit for visualization and inspection of multilingual models should be extremely valuable and Reference [83] is a good start in this direction.

Controlled Parameter Sharing. In between the extremities of parameter sharing exemplified by the above-mentioned models lies an array of choices. The degree of parameter sharing can be controlled at various layers of the MNMT system. A major factor driving the degree of parameter sharing is the divergence between the languages involved [123] and the desired simplicity as well as the flexibility of modeling.

Sharing encoders among multiple languages is very effective and is widely used [93, 123]. Keeping decoders separate is important, because the burden of generation is mainly on the decoder. Therefore, the job of the encoder is relatively simpler, which means that sharing an encoder between many languages leads to better parameter utilization. However, the decoder and its attention mechanism should be as robust as possible. Blackwood et al. [13] explored target language, source language, and pair-specific attention parameters. They showed that target language-specific attention performs better than other attention-sharing configurations, thus highlighting that designing a strong decoder is extremely important. For self-attention-based NMT models, Sachan and Neubig [123] explored various parameter-sharing strategies. They showed that sharing the decoder self-attention and encoder-decoder cross-attention parameters is useful for linguistically dissimilar languages. By sharing self- and cross-attention mechanisms in the decoder, the decoder most likely learns target language representations that are better aligned with source language representations. Wang et al. [157] further proposed a mechanism to generate a universal representation instead of separate encoders and decoders to maximize parameter sharing. They also used language-sensitive embedding, attention, and discriminator for different languages. This helps control the amount of sharing in an indirect fashion. Bapna and Firat [11] also extend a fully shared model with language-pair-specific adaptor layers that are fine-tuned for those pairs. After training a fully

[5]In particular, they focused on the effects of temperature-based data sampling on the multilingual models.

shared model, additional adaptor layers are inserted into the model and only those layers are fine-tuned, which requires significantly less computation cost. This does lead to an increase in the number of parameters but it is modest compared to a minimally shared system in which we would have to introduce many more parameters due to the addition of more encoders and decoders.

Fixing sharing configurations prior to training is ill-advised, because sharing one set of parameters might be optimal for one language pair but not another. To this end, Zaremoodi et al. [164] proposed a routing network to dynamically control parameter sharing where the parts to be shared depend on the parallel corpora used for training. However, Platanios et al. [115] learned the degree of parameter sharing from the training data. This is achieved by defining the language-specific model parameters as a function of global parameters and language embeddings. A base set of parameters is transformed into another set of parameters for a specific language using linear projections. If θ is the set of base parameters, then the parameters for a specific language pair src_i and tgt_j can be obtained via a linear transformation $F(\theta)$. This linear transformation involves learning some additional parameters for projection, but this involves far fewer parameters compared to modeling complete sets of parameters for each new language and thus this approach is quite attractive. It will be interesting to determine whether using non-linear projections is better than linear ones. This approach reduces the number of language-specific parameters (only language embeddings), while still allowing each language to have its own unique parameters for different network layers. In fact, the number of parameters is only a small multiple of the compact model (the multiplication factor accounts for the language embedding size) [70], but the language embeddings can directly impact the model parameters instead of the weak influence that language tags have.

Designing the right sharing strategy is important to maintaining a balance between model compactness, translation accuracy, simplicity, and flexibility. There are not many works that address the representation bottlenecks in these types of models. However, we should note that work by Huang et al. [65] focuses on efficient training mechanisms for extremely large multilingual models with over 128 layers so they can learn high-quality representations. Although, they do not actually analyze the learned representations it could be beneficial to analyze such models to understand how representation bottlenecks can be overcome. It will be interesting to see more approaches that use the training data itself to enable a model to increase or decrease its own complexity or capacity. Reinforcement learning and genetic algorithms applied to neural architecture search (NAS) [169] can be one of the ways to achieve this. We also refer the readers to work on conditional computation [12] in which parts of the model to be computed are determined by the data.

3.2 Addressing Language Divergence

A central task in MNMT is alignment of representations of words and sentences across languages so divergence between languages can be bridged, enabling the model to handle many languages. This involves the study and understanding of the representations learned by multilingual models and using this understanding to further improve modeling choices. The remainder of this subsection discusses these issues related to multilingual representations.

Vocabulary. An important but not substantially addressed aspect of MNMT (NMT in general) is vocabulary. In the case of MNMT models that share vocabularies, the most common way to create a vocabulary is to sample an equal number of words for each language to ensure balanced representation. These vocabularies can then be used to learn sub-word vocabularies that can eliminate the out-of-vocabulary issue for NMT. However, this does not account for the nature of the languages involved and the skew in their sizes. Recent works attempt to address this using the concept of temperature based vocabulary sampling [2] and this is known to have a positive impact on MNMT

models. This work was conducted in a massively multilingual setting, and we believe that further studies and innovation are required.

The Nature of Multilingual Representations. Since MNMT systems share network components across languages, they induce a relationship among representations of sentences across languages. Understanding the nature of these multilingual representations can help get insights into the working of multilingual models. Some works that visualize multilingual model embeddings suggest that the encoder learns similar representations for similar sentences across languages [36, 70]. Since these visualizations are done in very low dimensions (2–3 dimensions), they might not convey the right picture about the language-invariance of multilingual representations. Kudugunta et al. [83] do a systematic study of representations generated from a massively multilingual system using SVCCA [118], a framework for comparing representations across different languages, models, and layers. Their study brings out the following observations on the nature of multilingual embeddings from a compact MNMT system:

- While encoder representations for similar sentences are similar across languages, there is a fine-grained clustering based on language similarity. This explains why transfer learning works better with related languages (as discussed in Section 4).
- The boundary between the encoder and decoder is blurry and the source language representations depend on the target language and viceversa.
- Representation similarity varies across layers. The invariance increases in higher layers on the encoder side. However, the invariance decreases in higher layers on the decoder side. This is expected, since the decoder is sensitive to the target language to be generated. The decoder has to achieve the right balance between language-agnostic and language-aware representations.

Language-invariant representations seem like a nice abstraction, drawing comparisons to interlingual representations [70]. It has been a major goal of most MNMT research, as discussed in the subsequent sections. They have been shown to be beneficial for building compact models and transfer learning. Given these empirical observations from multiple sources, we think that language-invariant representations in appropriate parts of the multilingual model is a desirable property.

Encoder Representation. There are two issues that might make encoder representations language-dependent. Parallel sentences, from different source languages, can have a different number of tokens. Hence, the decoder's attention mechanism sees a variable number of encoder representations for equivalent sentences across languages. To overcome this, an attention bridge network generates a fixed number of contextual representations that are input to the attention network [94, 149]. By minimizing the diversity of representations, the decoder's task is simplified and it becomes better at language generation. The choice of a single encoder for all languages is also promoted by Hokamp et al. [64], who opt for language-specific decoders. Murthy et al. [101] pointed out that the sentence representations generated by the encoder are dependent on the word order of the language and are, hence, language-specific. They focused on reordering input sentences to reduce the divergence caused due to different word orders to improve the quality of transfer learning.

Decoder Representation. The divergence in decoder representation needs to be addressed when multiple target languages are involved. This is a challenging scenario, because a decoder should generate representations that help it generate meaningful and fluent sentences in each target language. Therefore, a balance between learning language-invariant representations and being

able to generate language-specific translations is crucial. If several parallel corpora for different language pairs are simply concatenated and then fed to a standard NMT model, then the NMT model might end up generating a mixed language translation as a result of vocabulary leakage due to language-invariant representations. The *language tag* trick has been very effective in preventing vocabulary leakage [70], because it enables the decoder to clearly distinguish between different languages.[6] Further, Blackwood et al. [13] added the language tag to the beginning as well as end of sequence to enhance its effect on the sentence representations learned by a left-to-right encoder. This shows that dedicating a few parameters to learn language tokens can help a decoder maintain a balance between language-agnostic and language-distinct features. Hokamp et al. [64] showed that more often than not, using separate decoders and attention mechanisms gives better results as compared to a shared decoder and attention mechanism. This work implies that the best way to handle language divergence would be to use a shared encoder for source languages and different decoders for target languages. We expect that the balance between language-agnostic and language-specific representations should depend on the language pairs. Prasanna [117], Tan et al. [141] are some of the works that cluster languages into language families and train separate MNMT models per family. Language families can be decided by using linguistic knowledge[7] [117] or by using embedding similarities where the embeddings are obtained from a multilingual word2vec model [141]. Comparing language-family-specific models and mixed-language-family models shows that the former models outperform the latter models. In the future, when training a model on a large number of languages, researchers could consider different decoders for different language families and each decoder in turn can use the language token trick to generate specific languages in the language family assigned to the decoder.

Impact of Language Tag. There are some works that explore the effects of the *language tag* trick on the shared decoder, its language-agnostic (or lack thereof) representations, and the final translation quality. Wang et al. [156] explored multiple methods for supporting multiple target languages: (a) target language tag at beginning of the decoder, (b) target language dependent positional embeddings, and (c) divide hidden units of each decoder layer into shared and language-dependent ones. Each of these methods provides gains over Johnson et al. [70], and combining all methods gave the best results. Hokamp et al. [64] showed that in a shared decoder setting, using a task-specific (language pair to be translated) embedding works better than using language tokens. We expect that this is because learning task-specific embeddings needs more parameters and helps the decoder learn better features to distinguish between tasks. In the related sub-area of pre-training using MASS [138], mBERT/XLM [32] it was shown that using language or task tags is sufficient for distinguishing between the languages used for pre-training large models. Given that massive pre-training is relatively new, further exploration is much needed.

In the above discussion, we have not considered the possibility of task interference and negative language transfer. Previous research has shown that certain languages help improve translation quality when grouped together [36, 170]. This implies the possibility of language pairs that may interfere with each other (task interference in the multi-task setting) and cause unhelpful knowledge to be transferred (negative transfer). Deeper research on this is pending.

[6]Our own experiments in extremely low-resource scenarios show that it is impossible to completely avoid vocabulary leakage, especially when the same word is present in different languages. Employing a special bias vector helps mitigate this issue.

[7]https://en.wikipedia.org/wiki/List_of_language_families.

3.3 Training Protocols

MNMT model training is a critical issue that requires sophisticated methods. The core of all methods is the minimization of the negative log-likelihood of the translation for all language pairs. Conceptually, the negative log-likelihood of the MNMT model is the average of the negative log-likelihoods of multiple bilingual models. As we have mentioned before, minimizing the negative log-likelihood can be done using one or more gradient-descent-based optimization algorithms such as SGD, ADAM, ADAGRAD, and so on [122]. There are two main types of training approaches: single stage or parallel or joint training and sequential or multi-stage training. Depending on the use-case, multi-stage training can be used for model compression (*knowledge distillation*) or fine-tuning, addition of data and/or languages (*incremental training*).

Single Stage Parallel/Joint Training. We simply pre-process and concatenate the parallel corpora for all language pairs and then feed them to the model batch-wise. For models with separate encoders and decoders, each batch consists of sentence pairs for a specific language pair [48]; whereas for fully shared models, a single batch can contain sentence pairs from multiple language pairs [70, 93]. As some language pairs might have more data than other languages, the model may be biased to translate these pairs better. To avoid this, sentence pairs from different language pairs are sampled to maintain a healthy balance. The most common way of maintaining balance is to oversample smaller datasets to match the sizes of the largest datasets, but Arivazhagan et al. [5] worked on temperature-based sampling and showed its cost-benefit analysis.

Knowledge Distillation. Knowledge distillation was originally proposed by Hinton et al. [63] for tasks that do not involve generating sequences such as image classification. The underlying idea is to train a large model with many layers and then distill its knowledge into a small model with fewer layers by training the small model on the softmax generated by the large model instead of the actual labels. This approach does not work well for sequence generation and, hence, Kim and Rush [73] proposed sequence distillation where they simply translate the training data using a pre-trained model. They then use this pseudo-data to train smaller models. These smaller models are able to learn faster and better because of the reduced burden for decision making. Following this, Tan et al. [142] trained bilingual models for all language pairs involved and then these bilingual models are used as *teacher models* to train a single *student model* for all language pairs. The student model is trained using a linear interpolation of the standard likelihood loss as well as distillation loss that captures the distance between the output distributions of the student and teacher models. The distillation loss is applied for a language pair only if the teacher model shows better translation accuracy than the student model on the validation set. This approach shows better results than joint training of a black-box model, but training time increases significantly, because bilingual models also have to be trained.

Incremental Training. These approaches aim to decrease the cost of incorporating new languages or data in multilingual models by avoiding expensive retraining. This is because not all language pairs may be available when the initial MNMT model is trained. It will be expensive to re-train the multilingual model from scratch when a language pair has to be added. A practical concern with training MNMT in an **incremental** fashion is dealing with vocabulary. Some works alter a pre-trained model's capacity to accommodate new languages. Lakew et al. [90] updated the vocabulary of the parent model with the low-resource language pair's vocabulary before transferring parameters. Embeddings of words that are common between the low- and high-resource languages are left untouched and randomly initialized embeddings may be used for as yet unseen words. A simpler solution would be to consider a universal romanized script so incorporating a new language can be a matter of resuming training or fine-tuning. Escolano et al. [45] focused

on first training bilingual models and then gradually increasing their capacity to include more languages. To address representation bottlenecks, Bapna and Firat [11] proposed expanding the capacities of pre-trained MNMT models (especially those trained on massive amounts of multilingual data) using tiny feed-forward components that they call adaptors. For each language pair in a pre-trained (multilingual) model, they add adaptors at each layer and fine-tune them on parallel corpora for that direction. These modular adaptor layers can be considered as experts that specialize for specific language pairs and can be added incrementally. They showed that this approach can help boost the performance of massively multilingual models trained by Arivazhagan et al. [5] and Aharoni et al. [2]. Note that this model can be used to incorporate new data, but new language pairs cannot be added. For multiway models, it is critical for incremental training to deal with catastrophic forgetting where previously learned knowledge is forgotten by including an additional language pair or training on additional data. While previously mentioned work on adaptor layers [11] is an effective solution, further research will be beneficial.

A major criticism of all these approaches is that the MNMT models are trained in the same way as a regular NMT model. Most researchers tend to treat all language pairs equally, with the exception of oversampling smaller corpora to match the sizes of the larger corpora, and tend to ignore the fact that NMT might be able to handle some language pairs better than others. There are works that propose to focus on scaling learning rates or gradients differently for high-resource and low-resource language pairs [67]. However, Kiperwasser and Ballesteros [74] proposed a multitask learning model for learning syntax and translation, where they showed different effects of their model for high-resource and low-resource language pairs. Furthermore, different language pairs might be learned at different rates and the optimal stopping criteria is also heavily underresearched. Currently, model selection is based on an automatic evaluation metric score on the dev set composed of multiple language pairs. As the metric is an average score for multiple language pairs, the chosen model might not be optimal for individual pairs. Therefore, we suppose that the MNMT models in the vast majority of papers are sub-optimally trained even if they improve the translation quality for low-resource languages. In hindsight, this particular aspect of MNMT training deserves more attention.

Although multiway MNMT modeling has been thoroughly explored, there are still a number of open questions, especially the representation bottleneck and parameter-sharing protocols. The next section will pay special attention to MNMT in resource-constrained scenarios.

4 MNMT FOR LOW-RESOURCE LANGUAGE PAIRS

Many language pairs have a limited amount of parallel corpora, which is insufficient for training high-quality NMT systems. While data augmentation strategies such as back-translation [130] and self-training [61] can improve translation quality via regularization and domain adaptation, parallel training signals are limited. Can MNMT provide additional parallel training signals from a high-resource language pair (e.g., Spanish-English) to improve low-resource MT (e.g., Catalan-English)? Since MNMT systems strive to represent multiple languages in the same vector space, it should be possible to utilize data from high-resource language pairs to improve translation of low-resource language pairs. Such *transfer learning* [112] approaches have been widely explored in the literature. The high-resource language (model) is often referred to as a *parent language (model)* whereas the low-resource language (model) is known as a *child language (model)*. The related literature has focused on four aspects of transfer learning: (1) training procedures to ensure maximum transfer, (2) addressing lexical divergence between the parent and child languages, (3) dealing with the differences in syntax between the parent and child languages, and (4) effect and utilization of language relatedness. The remainder of this section discusses these aspects as depicted in Figure 5.

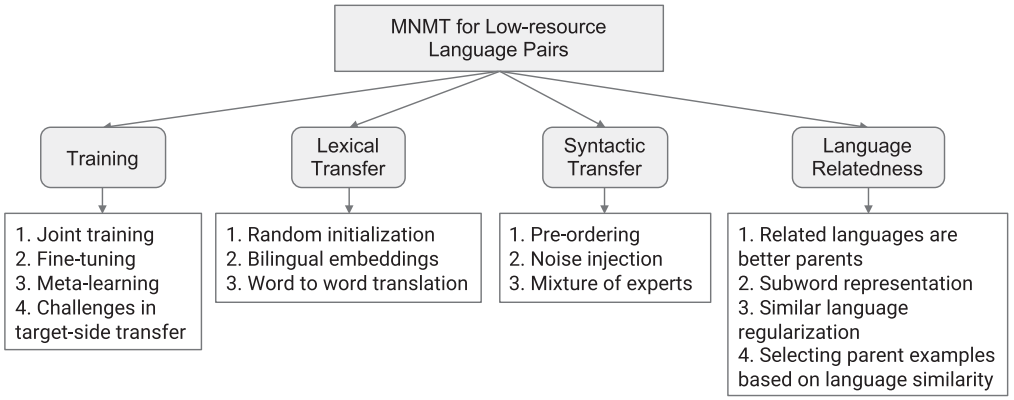

Fig. 5. An overview of MNMT for low-resource languages pairs depicting various aspects of proposed solutions.

4.1 Training

Most studies have explored transfer learning on the source-side: The high-resource and low-resource language pairs share the same target language. The simplest approach is **jointly training** both language pairs [70]. However, the final model may not be optimally tuned for the child language pair, whose performance we are interested in.

A better approach is to **fine-tune** the parent model with data from the child language pair. Zoph et al. [170] showed that such transfer learning can benefit low-resource language pairs. First, they trained a *parent model* on a high-resource language pair. The *child model* is initialized with the parent's parameters wherever possible and trained on the small parallel corpus for the low-resource pair. They also studied the effect of fine-tuning only a subset of the child model's parameters (source and target embeddings, RNN layers, and attention). They observed that fine-tuning all parameters except the input and output embeddings was optimal in one setting. However, further experiments are required for drawing strong conclusions.

Training the parent model to optimality may not be the best objective for child tasks. It may be preferable that parent model parameters are amenable to fast adaptation/fine-tuning on child tasks. Learning such a parent model is referred to as **meta-learning** and can generalize well to child tasks. Gu et al. [55] used the model-agnostic meta-learning (MAML) framework [47] to learn appropriate parameter initialization from the parent pair(s) by taking the child pair into consideration. The meta-learning-based approach significantly outperforms simple fine-tuning. They also show that having more parent language pairs (jointly trained) also improves performance on the child language pairs.

Transfer learning on the **target-side has been more challenging** than transfer learning on the source-side. Distinct target languages require target-language-specific representations, while transfer learning prefers target-language-invariant representations. The success of transfer learning relies on achieving the right balance between these factors. Johnson et al. [70] showed that joint training does not provide any significant benefit. Fine-tuning is beneficial in very low-resource scenarios [37], but gains may be limited due to catastrophic forgetting. Dabre et al. [37] showed that a multi-stage fine-tuning process is beneficial when multiple target languages are involved. They do not focus on language divergence during their multilingual multi-stage tuning but show that the size of helping data matters. From their work, it will be worthwhile to consider involving multiple medium-sized (few hundreds of thousands of lines) helping corpora involving a variety of languages. This multilingual multi-stage transfer learning scenario requires further research.

4.2 Lexical Transfer

Zoph et al. [170] **randomly initialized the word embeddings** of the child source language, because those could not be transferred from the parent. However, this approach does not map the embeddings of similar words across the source languages *a priori*. Gu et al. [54] improved on this simple initialization by **mapping pre-trained monolingual word embeddings** of the parent and child sources to a common vector space. This mapping is learned via the orthogonal Procrustes method [125] using bilingual dictionaries between the sources and the target language [61]. Kim et al. [71] proposed a variant of this approach where the parent model is first trained and monolingual word-embeddings of the child source are mapped to the parent source's embeddings prior to fine-tuning. While Gu et al. [54] require the child and parent sources to be mapped while training the parent model, the mapping in Kim et al. [71]'s model can be trained after the parent model has been trained.

4.3 Syntactic Transfer

It is not always possible to have parent and child languages from the same language family and, hence, blindly fine-tuning a parent model on the child language data might not take into account the syntactic divergence between them. Although it is important to address this issue there are surprisingly few works that address it. Murthy et al. [101] showed that reducing the *word order divergence* between source languages by **reordering** the parent sentences to match child word order is beneficial in extremely low-resource scenarios. Since reordering is part of the pre-processing pipeline, it is referred to as *pre-ordering*. Pre-ordering has been widely used in phrase-based MT [31] and explored in NMT [165] to overcome source-target syntactic divergence. Pre-ordering can be done via hand-written rules [87, 120] or learnt rules/models [151, 162]. We expect that future work will focus on syntactic divergence between languages via NMT models that handle reordering implicitly rather than rely on pre-ordering. Kim et al. [71] took a different approach to mitigate syntactic divergence. They trained the parent encoder with **noisy source data** introduced via probabilistic insertion and deletion of words as well as permutation of word pairs. This ensures that the encoder is not over-optimized for the parent source language syntax. Gu et al. [54] proposed to achieve better transfer of syntax-sensitive contextual representations from parents using **a mixture of language experts network**.

4.4 Language Relatedness

Results from the above-mentioned approaches show that, in many cases, transfer learning benefits the child language pair even if the parent and child languages are not related. These benefits are also seen even when the languages have different writing systems. These are interesting findings and can be attributed to the ability of the NMT models to learn cross-lingual representations, as discussed in previous sections. It is natural to ask if language relatedness can impact the effectiveness of lexical and syntactic transfer. Learning cross-lingual embeddings, used for lexical transfer, is difficult for distant languages [114, 137].

Zoph et al. [170] and Dabre et al. [40] empirically showed that **a related parent language benefits the child language more than an unrelated parent**. Maimaiti et al. [95] further showed that using multiple highly related high-resource language pairs and applying fine-tuning in multiple rounds can improve translation performance more, compared to only using one high-resource language pair for transfer learning. Kocmi and Bojar [76] presented a contradictory result in the case of Estonian (related Finnish vs. unrelated Czech/Russian parent) and suggest that size of the parent is more important. Another factor that may have aided transfer using Russian/Czech is their contact relationship with Estonian, since Russian and Czech are spoken in geographically

nearby regions to Estonia. Most of the literature suggests that language relatedness is beneficial. It is probably easier to overcome language divergence when the languages are related, but a deeper study of how to uncover different factors is warranted.

Further, language relatedness can be explicitly utilized to improve transfer learning. Language relatedness is typically exploited by using shared **subword BPE vocabulary and BPE embeddings** between the parent and child languages [106]. Maimaiti et al. [95] used a unified transliteration scheme at the character level. This utilizes the **lexical similarity** between the languages and shows significant improvements in translation quality. Neubig and Hu [105] used **"similar language regularization"** to prevent overfitting when rapidly adapting a pre-trained, massively multilingual NMT model (universal model) for low-resource languages. While fine-tuning the universal model for a low-resource pair, overfitting is avoided by using a subset of the training data for a related high-resource pair along with the low-resource pair. Chaudhary et al. [19] used this approach to translate 1,095 languages to English. Further, not all parallel data from the parent task may be useful in improving the child task. Wang and Neubig [154] proposed **selection of sentence pairs from the parent task** based on the similarity of the parent's source sentences to the child's source sentences. The significant gains from simple methods described point to the value of utilizing language relatedness. Further methods should be explored to create language-invariant representations specifically designed for related languages.

While low-resource translation is hard in itself, an even more extreme scenario is where no direct data exists between language pairs of interest. The next section discusses literature related to this scenario.

5 MNMT FOR UNSEEN LANGUAGE PAIRS

Providing machine translation solutions for arbitrary language pairs remains a challenge, since little to no parallel corpora exist for most language pairs.[8] Unsupervised NMT [92] has shown promise and generated some interest in recent years, but their quality remains way behind supervised NMT systems for most language pairs.

Can we do better than unsupervised NMT by utilizing multilingual translation corpora? A key observation is: Even if two languages do not have a parallel corpus, they are likely to share a parallel corpus with a third language (called the *pivot* language). In many cases, English is likely to be the pivot language, given its widespread global usage. Pivot-language-based translation is a type of multilingual MT involving corpora for two language pairs: source-pivot and pivot-target parallel corpora, and has been widely explored for translation between unseen languages. In addition to simple pivot translation, zero-shot and zero-resource MNMT approaches have been proposed for translation between unseen language pairs. These approaches are described in the reminder of this section. Figure 6 gives an overview of the major approaches and issues in this regard.

5.1 Pivot Translation

The simplest approach to pivot translation is building independent source-pivot (S-P) and pivot-target (P-T) MT systems. At test time, the source sentence cascades through the S-P and P-T systems to generate the target sentence. Note that this pivot method is agnostic to the underlying translation technology and **can be applied to SMT [147], RBMT [160], or NMT [91] systems**. Also, the pivoting can be done using an MNMT system, too [56, 91] (as opposed to pivoting via bilingual systems). This simple process has two **limitations due to its pipeline characteristic**: (a) translation errors compound in a pipeline; and (b) decoding time is doubled, since inference

[8]N-way translations of the Bible may be amongst the rare source of parallel corpora across arbitrary language pairs, but it is not sufficient for training a general-purpose MT system.

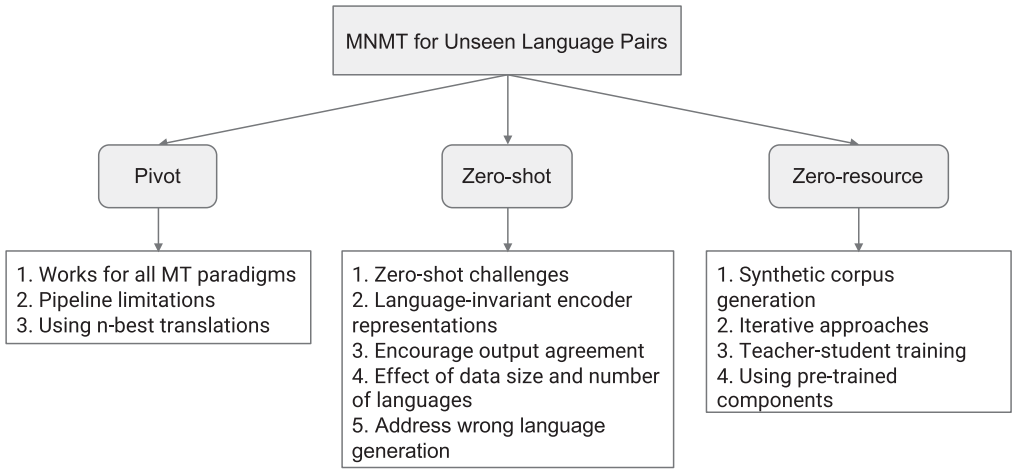

Fig. 6. An overview of MNMT for unseen languages pairs. There are three broad approaches: pivot, zero-shot, and zero-resource approaches.

has to be run twice. Sometimes, more than one pivot may be required to translate between the source and pivot language. Increased pipeline length exacerbates the above-mentioned problems. The quality of the source-pivot translation is a bottleneck to the system. A variant of this approach extracts **n-best translations** from the S-P system. For each pivot translation, the P-T system can generate m-best translations. The $n \times m$ translation candidates can be re-ranked using scores from both systems and external features to generate the final translation. This approach improves the robustness of the pipeline to translation errors.

5.2 Zero-shot Translation

Multilingual NMT models offer an appealing possibility. Even if the MNMT system has not been trained for the unseen language pair, Johnson et al. [70] showed that the system is able to generate reasonable target-language translations for the source sentence. Along with the source sentence, the desired output language's language tag is provided as input. This is sufficient for the system to generate output in the target language. Note that the MNMT system was exposed to zero bilingual resources between the source and target languages during training and encountered the unseen language pair only at test-time. Hence, this translation scenario is referred to as *zero-shot translation*.

The appeal of zero-shot translation is two-fold:

- Translation between any arbitrary language pair can be done in a single decoding step, without any pivoting, irrespective of the number of implicit pivots required to bridge the source and target language in the multilingual training corpus. The MNMT system can be conceived as an *implicit pivoting* system.
- Given a multilingual parallel corpus spanning N languages, only a single multilingual NMT model is required to translate between $N \times (N-1)$ languages.

Challenges of Zero-shot Translation. It has been argued that training an MNMT system incorporating multiple languages could benefit zero-shot translation due to better interlingual representations and elimination of cascading errors [70]. The simple zero-shot system described above, though promising, belies these expectations and its performance is generally lower than the pivot translation system [4, 56, 70, 113].

Some researchers have analyzed zero-shot translation to understand its underperformance. The following reasons have been suggested as limitations of zero-shot translation:

Spurious correlations between input and output language. During training, the network is not exposed to the unseen pairs. In its quest to capture all correlations in the training data, the model learns associations between the input representations and the target language for the observed language pairs. At test time, irrespective of the target language tag, the network will tend to output a language it has already observed with the source language in the training data [56]. Arivazhagan et al. [4] showed that translation quality is closer to pivot systems if evaluation is restricted to sentences where the correct target language is generated.

Language variant encoder representations. The encoder representations generated by the model for equivalent source and pivot languages are not similar. Due to this discrepancy, the output generated by the decoder will be different for the source and pivot representations [4, 83].

To address these limitations, the following methods have been proposed for improving zero-shot translation:

Minimize divergence between encoder representations. During training, additional objectives ensure that the source and pivot encoder representations are similar. Arivazhagan et al. [4] suggested an unsupervised approach to align the source and pivot vector spaces by minimizing a domain adversarial loss [50]—a discriminator is trained to distinguish between different encoder languages using representations from an adversarial encoder. Since S-P parallel corpora is also available, supervised loss terms that penalize divergence in source and pivot representations for each sentence pair have also be explored. Different loss functions such as cosine distance [4], Euclidean distance [113], and correlation distance [124] have been shown to be beneficial in reducing the source/pivot divergence. Ji et al. [69] proposed to use pre-trained cross-lingual encoders trained using multilingual MLM, XLM, and BRLM objectives to obtain language-invariant encoder representations. Sen et al. [129] used denoising autoencoding and back-translation to obtain language-invariant encoder representations.

Encourage output agreement. Al-Shedivat and Parikh [3] incorporated additional terms in the training objective to encourage source and pivot representations of parallel sentences to generate similar output sentences (synthetic) in an auxiliary language (possibly an unseen pair). This also avoids spurious correlations, since the network learns to generate unseen pairs. Similar considerations motivated Pham et al. [113] to add a pivot auto-encoding task in addition to the source-pivot translation task. They incorporate additional loss terms that encourage the attention-context vectors as well as decoder output representations to agree while generating the same pivot output. Xu et al. [163] considered different translation paths among multiple languages in unsupervised NMT by designing training objectives for these paths to achieve the same goal.

Effect of corpus size and number of languages. Aharoni et al. [2] suggested that the zero-shot performance of multilingual NMT system increases with the number of languages incorporated in the model. It is not clear if the approaches mentioned above to address zero-shot NMT limitations can scale to a large number of languages. Arivazhagan et al. [4] showed that cosine distance-based alignment can be scaled to a small set of languages. Some studies suggest that zero-shot translation works reasonably well only when the multilingual parallel corpora is large [91, 97].

Addressing wrong language generation. To address the problem of generation of words in the wrong language, Ha et al. [59] proposed to filter the output of the softmax, forcing the model to translate into the desired language. This method is rather effective despite its simplicity.

The zero-shot approaches discussed above can complement multiway NMT systems described in Section 3 to support translation between the language pairs the model has not observed during

training. Further work is needed to establish if these methods can scale to massively multilingual models.

5.3 Zero-resource Translation

The zero-shot approaches discussed in the previous section seek to avoid any training specific to unseen language pairs. Their goal is to enable training massively multilingual NMT systems that can perform reasonably well in zero-shot directions, without incurring any substantial overhead during training for all unseen language heads. When there is a case for optimizing the translation quality of an unseen pair, the training process can also consider objectives specific to the language pair of interest or tune the system specifically for the language pair of interest. Such approaches are referred to as *zero-resource* approaches [49]. Their training objectives and regimen customized for the unseen language pair of interest distinguish them from *zero-shot* approaches. Note that zero-resource approaches do not use any true source-target parallel corpus. The following approaches have been explored for zero-resource translation.

Synthetic Corpus Generation. The pivot side of the P-T parallel corpus is back-translated to the source language. The back-translation can be achieved either through zero-shot translation or pivot translation, creating a synthetic S-T parallel corpus. An S-T translation model can be trained on this synthetic parallel corpus. Adding the synthetic corpus to the multilingual corpus helps alleviate the spurious correlation problem. Some works have shown that this approach can outperform the pivot translation approach [49, 56, 91]. The source is synthetic, hence, there is a difference between the training and test scenarios. Further, synthetic parallel corpus can be generated from monolingual pivot data, too [35].

Iterative Approaches. The S-T and T-S systems can be trained iteratively such that the two directions reinforce each other [91]. Sestorain et al. [132] jointly trained both the models incorporating language modelling and reconstruction objectives via reinforcement learning. The LM objective ensures grammatical correctness of translated sentences, while the reconstruction objective ensures translation quality. The major shortcoming of iterative approaches is that they usually do not yield improvements after the first 2–3 iterations and are extremely time-consuming.

Teacher-student Training. Chen et al. [21] assumed that the source and pivot sentences of the S-P parallel corpus will generate similar probability distributions for translating into a third language (target). They build the S-T model ("student") without any parallel S-T corpus by training the model to follow a P-T model ("teacher"). They propose two approaches: sentence-level mode approximation and word-level KL divergence. Both approaches outperform the pivot baseline, with the latter showing better performance. A shortcoming of the first method is reliance on S-T parallel corpus with synthetic target, whereas the second approach learns to follow a soft distribution.

Combining Pre-trained Encoders and Decoders. Kim et al. [72] combined S-P encoder with P-T decoder to create the S-T model. They improved the simple initialization using some fine-tuning objectives and/or source-pivot adaptation to ensure that source and pivot representations are aligned. Whether this can be done via a fully shared MNMT model or not is unclear.

Most the methods discussed in this section are either just competitive with or slightly better than simple pivot translation. This opens opportunities in understanding the challenges in unseen language translation and exploring solutions.

6 MULTI-SOURCE NMT

If a source sentence has already been translated into multiple languages, then these sentences can be used together to improve the translation into the target language. This technique is known

Table 1. Overview of Multi-source NMT Approaches Based on the Availability of N-way Data, the Training Concerns, and Benefits Concerns

Multi-Source Approach		N-way data needed	Solutions	Concerns	Benefits
1	**Vanilla**	Yes	Multi or Shared Encoder model	Long training times; Bulky Models	Expanding N-way corpora
2	**Ensembling**	No	Ensemble multiple bilingual models	Learning ensemble functions; Need small N-way corpora	Reuse existing bilingual models
3	**Synthetic data**	No	Generate missing source sentences	Infeasible for real-time translation	Applicable for post-editing

as multi-source MT [110]. The underlying principle is to leverage complementarity in terms of source-side linguistic phenomena expressed in multiple languages.

Why Multi-source MT? At first, it may seem unrealistic to assume the availability of the same sentence in multiple languages, but we would like to point to the parliamentary proceedings in the European Union (EU) or India. In the EU there are more than 10 official languages and in India there are more than 22 official languages. The EU parliamentary proceedings are maintained in multiple languages especially when they are concerning issues affecting multiple member nations. Therefore, it is extremely common for human interpreters to simultaneously interpret a speaker's language into several other languages. In such a situation, instead of employing a large number of interpreters for each language, it is possible to employ interpreters for a subset of languages and then leverage multi-source MT of these subsets of languages, which can give much better translation quality as compared to singe-source MT. Furthermore, these improved translations can be edited by translators and added to the dataset that is used to train multi-source MT systems. Thus, it will be possible to create N-lingual (N > 3) corpora such as Europarl [77] and UN [167]. Refer to Table 1 for a simplified overview of the multi-source NMT paradigm. There are three possible resource/application scenarios where multi-source NMT can be used.

Multi-source Available. Most studies assume that the same sentence is available in multiple languages, although this need not be true. However, whenever multi-source sentences are available, it is crucial to leverage them. Just like multiway NMT models, a multi-source NMT model can consist of multiple encoders or a single encoder. Zoph and Knight [168] showed that a multi-source NMT model using *separate encoders and attention networks* for each source language outperforms single-source models. In their model, each encoder generates representations for different source-language sentences and the decoder has a separate attention mechanism per source language, and the attention weight in Equation (2) is calculated as:

$$e_{ji}^{src_l} = a\left(s_{j-1}, h_i^{src_l}\right),$$

where src_l indicates a source language. The separate attentions are concatenated and then used in the decoder. This leads to an increase in the hidden layer size of the decoder and, hence, the number of decoder parameters. This is often acceptable unless large number of source languages are to be used. Although not explored anywhere, we suppose that a linear transformation can be used to down-project the concatenation of the context vectors and thus prevent the large hidden sizes in the decoder. Firat et al. [49], instead of training a special multi-source model, used a previously trained MNMT model. They generated the contexts for different source language as Zoph and Knight [168], but instead of concatenating and projecting them, they simply added them and fed this representation to the decoder. This process is known as early averaging. They also proposed a method known as late averaging, where for each time step in the decoder, the softmax is computed

using each individual source and then averaged. A combination of late and early averaging was shown to give the best results.

A simpler approach concatenated multiple *source sentences* and fed them to a standard NMT model [36], with performance comparable to Zoph and Knight [168]. This model was obviously inspired by the success of fully shared MNMT models [70]. A single encoder is responsible for encoding a long multilingual sentence.[9] Interestingly, this model could automatically identify the boundaries between different source languages and simplify the training process for multi-source NMT. Dabre et al. [36] also showed that it is better to use linguistically similar source languages, especially in low-resource scenarios. Both studies showed that the attention mechanisms tend to prefer some languages over others. Especially, linguistically distant languages are practically ignored with computing context for decoding. It is possible to speed up encoding by parallellizing the encoders.

Ensembling of individual source-target models is another beneficial approach, for which Garmash and Monz [51] proposed several methods with different degrees of parameterization. They proposed to learn an ensembling function to combine the softmaxes of several bilingual models. This approach requires a smaller N-lingual corpus, but training an ensembling function can be costly in terms of parameters. However, Dabre et al. [36] have shown that it is possible to perform naive ensembling[10] and still get reasonable improvements in translation quality. Note that the ensembling method used by Dabre et al. [36] is the same as the late averaging technique proposed by Firat et al. [49]. Although the black-box method by Dabre et al. [36] is significantly simpler and more elegant than the method in Reference [168], most works on multi-source NMT use the latter method.

Missing Source Sentences. There can be missing source sentences in multi-source corpora and during translation. Nishimura et al. [108] extended Reference [168] by representing each "missing" source language with a *dummy token*. The NMT model manages to adapt to missing sentences and thus manages to give improved translations when all source sentences are available. Choi et al. [25] and Nishimura et al. [107] proposed to use MT-generated *synthetic sentences*, instead of a dummy token for the missing source languages. NMT models are used to translate sentences from the available source languages into the missing ones. When sentences for all source languages are ready, any standard multi-source approach may be used. The fact that such a simple "trick" works so well means that combining all these methods allows the training of a single model with both single as well as multi-source capabilities. Therefore, a single model can be used in different deployment conditions. If low-latency translation services are desired, then the model can be used in single-source mode. In moderate-latency scenarios, only a few source languages could be used. In offline translation scenarios, all source languages could be used. Future work could focus on dynamically determining which source languages are useful and which are not. Furthermore, a multiway, multi-source model might be an interesting challenge to tackle.

Post-editing. Instead of having a translator translate from scratch, multi-source NMT can be used to generate high-quality translations. The translations can then be post-edited, a process that is less labor-intensive and cheaper compared to translating from scratch. Multi-source NMT has been used for post-editing where the translated sentence is used as an additional source, leading to improvements [18]. Multi-source NMT has also been used for system combination, which combines NMT and SMT outputs to improve translation performance [166].

[9]Note that the order of the input sentences should be the same during training and testing.
[10]This means uniform avergaging of softmaxes of different models as opposed to weighted averaging.

In general, multi-source NMT does not receive much attention, because it cannot be used for real-time translation, as it involves additional intermediate translation steps. However, it should be possible to develop a system that can flexibly perform multi-source, multi-target as well as single source and single target translation. The system by Firat et al. [49] is such a system, but the adaptability to language pairs other than European languages and multiple (more than two) source languages has not been verified. Such a model can be flexibly used during online as well as offline translation scenarios.

7 MULTILINGUALISM IN OLDER PARADIGMS

One of the long-term goals of the MT community has been the development of architectures that can handle more than two languages. There have been approaches to incorporate multilingualism in the older rule-based and classical SMT paradigms. This section gives a brief overview of these attempts and compares them with MNMT.

7.1 Rule-based Machine Translation

Although rule-based systems (RBMT) do not attract research interest, we have included this short discussion for completeness with a goal to understanding the connections of MNMT with previous approaches. RBMT systems using an *interlingua* were explored widely in the past. The interlingua is a symbolic semantic, language-independent representation for natural language text [133]. Two commonly used interlinguas are UNL [146] and AMR [9]. Different interlinguas have been proposed in various systems, such as KANT [109], UNL, UNITRAN [43], and DLT [158]. Language-specific analyzers converted language input to interlingua representation, while language-specific decoders converted the interlingua representation into another language. To achieve an unambiguous semantic representation, a lot of linguistic analysis had to be performed and many linguistic resources were required. Hence, in practice, most interlingua systems were limited to research systems or translation in specific domains and could not scale to many languages.

7.2 Statistical Machine Translation

Phrase-based SMT (PBSMT) systems [81], a very successful MT paradigm, were also bilingual for the most part. Compared to RBMT, PBSMT requires less linguistic resources and instead requires parallel corpora. However, like RBMT, they work with symbolic, discrete representations, making multilingual representation difficult. Moreover, the central unit in PBSMT is the *phrase*, an ordered sequence of words (not in the linguistic sense). Given its arbitrary structure, it is not clear how to build a common symbolic representation for phrases across languages. Nevertheless, some shallow forms of multilingualism have been explored in the context of: (a) pivot-based SMT, (b) multi-source PBSMT, and (c) SMT involving related languages.

Pivoting. Most commonly used solutions are: chaining source-pivot and pivot-target systems at decoding [147], training a source-target system using synthetic data generated using target-pivot and pivot-source systems [53], and phrase-table triangulation pivoting source-pivot and pivot-target phrase tables [147, 159].

Multi-source. Typical approaches are: re-ranking outputs from independent source-target systems [110], composing a new output from independent source-target outputs [98], and translating a combined input representation of multiple sources using lattice networks over multiple phrase tables [126].

Related Languages. For multilingual translation with multiple related source languages, the typical approaches involved script unification by mapping to a common script such as Devanagari

[10] or transliteration [103]. Lexical similarity was utilized using subword-level translation models [85, 86, 144, 150]. Combining subword-level representation and pivoting for translation among related languages has been explored [62, 88, 144]. Most of the above-mentioned multilingual systems involved either decoding-time operations, chaining black-box systems, or composing new phrase-tables from existing ones.

Comparison with MNMT. While symbolic representations constrain a unified multilingual representation, distributed universal language representation using real-valued vector spaces makes multilingualism easier to implement in NMT. As no language-specific feature engineering is required for NMT, it is possible to easily scale to multiple languages. Neural networks provide flexibility in experimenting with a wide variety of architectures, while advances in optimization techniques and availability of deep learning toolkits make prototyping faster.

8 DATASETS AND RESOURCES

MNMT requires parallel corpora in similar domains across multiple languages. The following publicly available datasets can be used for research.

Multiway. Commonly used publicly available multilingual parallel corpora are the TED corpus [99] (the TED corpus is also available in speech [41]), UN Corpus [167], and those from the European Union, such as Europarl, JRC-Aquis, DGT-Aquis, DGT-TM, ECDC-TM, EAC-TM [139]. While these sources are primarily composed of European languages, parallel corpora for some Asian languages is accessible through the WAT shared task [102]. Only a small amount of parallel corpora are available for many languages, primarily from movie subtitles and software localization strings [145]. Recently, the large-scale WikiMatrix corpus [128] and the JW300 corpus covering 300 low-resource languages [1] have been released. For Indic languages, the CVIT-PIB corpus [136], the PMIndia corpus [60], and the IndoWordNet parallel corpus [84] have also been released recently. Readers can refer to catalogs such as OPUS[11] and the IndicNLP catalog[12] for comprehensive listings of parallel corpora resources.

Low- or Zero-resource. For low- or zero-resource NMT translation tasks, good test sets are required for evaluating translation quality. The above-mentioned multilingual parallel corpora can be a source for such test sets. In addition, there are other small parallel datasets, such as the FLORES dataset for English-{Nepali,Sinhala} [57], the XNLI test set spanning 15 languages [34], *CVIT-Mann ki Baat* [136], and the Indic parallel corpus [116]. The WMT shared tasks [15] also provide test sets for some low-resource language pairs.

Multi-source. The corpora for multi-source NMT have to be aligned across languages. Multi-source corpora can be extracted from some of the above-mentioned sources. The following are widely used for evaluation in the literature: Europarl [77], TED [145], UN [167]. The Indian Language Corpora Initiative (ILCI) corpus [68] is an 11-way parallel corpus of Indian languages along with English. The Asian Language Treebank [143] is a 9-way parallel corpus of South-East Asian languages along with English, Japanese, and Bengali. The MMCR4NLP project [39] compiles language family grouped multi-source corpora and provides standard splits. The Bible corpus [26] is a massive N-way one originally containing 100 languages. It has further been translated in more than 1,000 languages and used for multilingual studies [96, 111].

Shared Tasks. Recently, shared tasks with a focus on multilingual translation have been conducted at IWSLT [16], WAT [102], and WMT [15]; so, common benchmarks are available.

[11] http://opus.nlpl.eu.
[12] https://github.com/indicnlpweb/indicnlp_catalog.

9 CONNECTIONS WITH DOMAIN ADAPTATION

High-quality parallel corpora are limited to specific domains. Both vanilla SMT and NMT perform poorly for domain-specific translation in low-resource scenarios [44, 80]. Leveraging out-of-domain parallel corpora and in-domain monolingual corpora for in-domain translation is known as domain adaptation for MT [30].

As we can treat each domain as a language, there are many similarities and common approaches between MNMT and domain adaptation for NMT. Therefore, similar to MNMT, when using out-of-domain parallel corpora for domain adaptation, multi-domain NMT and transfer-learning-based approaches [29] have been proposed for domain adaptation. When using in-domain monolingual corpora, a typical way of doing domain adaptation is generating a pseudo-parallel corpus by back-translating target in-domain monolingual corpora [130], which is similar to the pseudo-parallel corpus generation in MNMT [49].

There are also many differences between MNMT and domain adaptation for NMT. While pivoting is a commonly used approach for MNMT [23], it is unsuitable for domain adaptation. As there are always vocabulary overlaps between different domains, there are no zero-shot translation [70] settings in domain adaptation. In addition, it is not uncommon to write domain-specific sentences in different styles, and so multi-source approaches [168] are not applicable either. However, data selection approaches in domain adaptation that select out-of-domain sentences that are similar to in-domain sentences [152] have not been applied to MNMT. In addition, instance weighting approaches [153] that interpolate in-domain and out-of-domain models have not been studied for MNMT. However, with the development of cross-lingual sentence embeddings, data selection and instance weighting approaches might be applicable for MNMT in the near future.

There are also studies trying to connect MNMT and domain adaptation. These studies focus on applying or improving fine-tuning for both MNMT and domain adaptation. Chu and Dabre [27] conducted a preliminary study for this topic, where they transfer knowledge from both multilingual and out-of-domain corpora to improve in-domain translation for a particular language pair via fine-tuning. Chu and Dabre [28] focused on training a single translation model for multiple domains by either learning domain-specialized hidden-state representations or predictor biases for each domain and incorporating multilingualism into the domain adaptation framework. Dabre et al. [37] applied multi-stage fine-tuning on multiway MNMT, which has been shown effective in domain adaptation by Reference [29]. Bapna and Firat [11] improved the scalability of fine-tuning for both MNMT and domain adaptation. Instead of fine-tuning the entire NMT system, they propose using adapter layers with relatively smaller number of parameters that are suitable for the target task. Given the success of connecting MNMT and domain adaptation with fine-tuning, we believe that there is potential to connect them with other approaches, given their similarities and differences discussed above.

10 FUTURE RESEARCH DIRECTIONS

While exciting advances have been made in MNMT in recent years, there are still many interesting directions for exploration. Refer to Table 2 for possible future avenues for MNMT research. The following research directions are by no means exhaustive and are to be considered as guidelines for researchers wishing to address specific important problems in MNMT. Throughout the article, we have identified three major challenges as follows: finding the balance between language-agnostic and language-aware representations, improving transfer learning, and developing a single NMT model for all possible language pairs. The following topics, starting from the most promising, should help to further tackle these aforementioned challenges.

Table 2. Overview of Future Research Directions and Issues for MNMT

Central Goal	Possible Directions/Issues/Approaches
Language Representation Learning	1. Balancing the sharing of representations between languages. 2. Handling language divergence. 3. Addressing code-switching and dialects. 4. Identifying language families computationally.
Leveraging Pre-trained Models	1. Pre-trained BERT/GPT/Transformer XL encoders and decoders. 2. Incorporating web-level knowledge into translation process. 3. Designing pre-training objectives for multilingualism. 4. Dealing with large model sizes. 5. Universal parent (pre-trained) models.
One Model for All Languages	1. A single model for all languages, domains, dialects, and code-switching. 2. Possible improvement from multi-modal knowledge. 3. A model to explain multilingualism. 4. Handling representation bottleneck.

Exploring Pre-trained Models. Pre-training embeddings, encoders, and decoders has been shown to be useful for NMT [119]. Most works rely on fine-tuning but do not address techniques to maximize the impact of transfer. Furthermore, how pre-training can be incorporated into different MNMT architectures is important as well. Recent advances in cross-lingual word [6, 17, 33, 66, 75, 100] and sentence embeddings[13] [7, 20, 34, 46] could provide directions for this line of investigation. Currently, transfer learning through unsupervised pre-training on extremely large corpora and unsupervised NMT is gaining momentum and we believe that investing in these two topics or a merger between them will yield powerful insights into ways to incorporate large amounts of knowledge into translation systems. Recently, this research topic has shown to have a lot of promise towards improving bilingual MT quality and should be beneficial for MNMT as well.

Unseen Language Pair Translation. Most work on unseen language pair translation has only addressed cases where the pivot language is related to or shares the same script with the source language. In many cases, the pivot language (mostly English) is unlikely to be related to the source and target languages, and this scenario requires further investigation (especially for zero-shot translation). Moreover, new approaches need to be explored to significantly improve over the simple pivot baseline.

Fast Multi-source NMT. While multi-source NMT gives good results, it is limited by the lack of multiway data during decoding. In a low-latency setting, the only acceptable solution would be a way to rapidly generate the missing source sentences from the available source language. We suppose that a form of simultaneous multi-target NMT solution might be a starting place for an effective technique. The main idea would be to generate additional sources simultaneously along with the target language and then at each decoding step the partially synthesized source sentences can be used to provide additional contexts.

Related Languages, Language Registers, and Dialects. Translation involving related languages, language registers, and dialects can be further explored given the existing evidence about the importance of language relatedness for improving MNMT quality. For dialects, currently the focus is on modern standard dialects of languages, but most people feel comfortable speaking in their local dialect and, hence, it would be valuable to transfer translation knowledge obtained for modern standard languages to their dialects. In our opinion, unsupervised MT methods using monolingual data should be extremely effective at handling dialect translation.

[13] https://engineering.fb.com/ai-research/laser-multilingual-sentence-embeddings/.

Code-mixed Language. Addressing intra-sentence multilingualism, i.e., code-mixed input and output, creoles, and pidgins, is an interesting research direction, as it is a variation of related-languages translation. The compact MNMT models can handle code-mixed input, but code-mixed output remains an open problem [70]. Code-mixed languages are gradually evolving into languages with their own unique grammar, and just like dialects, a vast majority of people feel more comfortable with code-mixing. For this, we believe that researchers should first focus on identifying code-mixing phenomena (either linguistically or statistically) and then design multilingual solutions to address them.

Visualization and Model Inspection. Most works focus on an extrinsic evaluation of multilingual models. While the visualization of the embeddings of NMT models in multilingual settings has revealed some interesting behaviors [70, 83], we feel that more researchers should pursue this. Furthermore, we believe that visualization that relies on down-projecting high-dimensional vectors to 2-D and 3-D might not reveal actual phenomenon and lead to certain biases that affect MNMT modeling. It will be interesting to have works that perform analyses of high-dimensional representations and reveal the impact of multilingualism.

Learning Effective Language Representations. A core question that still needs further investigation is: What is the degree of language awareness that needs to be considered to learn task-effective continuous space representations? Particularly, the questions of word-order divergence between the source languages and variable length encoder representations have received little attention. Addressing this issue will eliminate the need to deal with language divergence issues for transfer learning. Most works tend to agree that language-agnostic encoders and language-aware decoders tend to work well but a deeper investigation is still needed.

Multiple Target Language MNMT. Most current efforts address multiple source languages. Multiway systems focusing on multiple low-resource target languages are important not only because it is a significantly harder challenge, but because it can help shed more light on language divergence and its impact on translation quality. The right balance between sharing representations vs. maintaining the distinctiveness of the target language for generation needs exploring. Furthermore, depending on the similarity between languages or lack thereof, it might be important to consider a hybrid architecture (mixing RNN, CNN, and FFN) for best performance.

Representation Bottleneck. Recent works on massively multilingual NMT have shown that existing approaches are unable to successfully incorporate more than 100 language pairs and at the same time improve translation quality, especially for the resource-rich languages (which tend to suffer). This is not a problem of the number of layers, encoders, or decoders in the NMT model but rather a problem of the way the MNMT model is currently designed. Current approaches that focus on either separate encoders or decoders lead to a large unwieldy model or shared encoders and decoders with a large number of parameters, both of which are hard to train. Works that leverage incremental learning or knowledge distillation seem to be promising.

Joint Multilingual and Multi-domain NMT. Jointly tackling multilingual and multi-domain translation is an interesting direction with many practical use cases. When extending an NMT system to a new language, the parallel corpus in the domain of interest may not be available. Transfer learning in this case has to span languages and domains. It might be worthwhile to explore adversarial approaches where domain and language-invariant representations can be learned for the best translations.

Multilingual Speech-to-speech NMT. Most work on MNMT, and NMT in general, has focussed on translating text. Recently, progress has been made in speech translation as well as in

multilingual ASR and TTS. An interesting research direction would be to explore multilingual speech translation, where the ASR, translation, and TTS modules can be multilingual. Interesting challenges and opportunities may arise in the quest to compose all these multilingual systems in an end-to-end method. Multilingual end-to-end speech-to-speech translation would also be a future challenging scenario. Some datasets are already available for such research [41]. We believe that this will be one of the hardest challenges in the field of multilingual translation.

11 CONCLUSION

MNMT has made rapid progress in recent years. In this survey, we have covered literature pertaining to the major scenarios we identified for multilingual NMT: multiway, low- or zero-resource (transfer learning, pivoting, and zero-shot approaches), and multi-source translation. We have systematically compiled the principal design approaches and their variants, central MNMT issues and their proposed solutions, along with their strengths and weaknesses. We have put MNMT in a historical perspective with respect to work on multilingual RBMT and SMT systems. At the outset, we raised two important research questions that we feel can be answered to a certain extent as below:

Q1. *Is it possible to have a one-model-for-all-languages solution to MT (NLP) applications?*
A1. *Yes. While it is possible to fit all language pairs into a single model, existing deep learning methodologies suffer from representation learning bottlenecks and generalization capabilities that put a limit on the gains from multilingualism on translation quality. Further research into better data selection and representation, network architectures, and learning algorithms is needed.*
Q2. *Can shared multilingual distributed representations help MT (NLP) for low-resource languages?*
A2. *Yes. Low-resource language translation improves significantly, but further research on balancing the language-agnostic and language-specific nature of representations can help push performance even further.*

It is clear that multilingualism can not only improve translation quality when leveraged but also can provide a number of insights into the relationships between languages. Most works lack human evaluation of multilingual models, which should help in better understanding the impact of multilingualism. We suggest promising and important directions for future work. We hope that this survey article will give researchers and practitioners a deeper understanding of the MNMT landscape and enable them to choose engaging topics for research. We hope that our work will help promote and accelerate MNMT research.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3204–3210. DOI: https://doi.org/10.18653/v1/P19-1310

[2] Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 3874–3884. Retrieved from https://www.aclweb.org/anthology/N19-1388.

[3] Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 1184–1197. Retrieved from https://www.aclweb.org/anthology/N19-1121.

[4] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *CoRR* abs/1903.07091 (2019).

[5] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR* abs/1907.05019 (2019).

[6] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2289–2294. DOI: https://doi.org/10.18653/v1/D16-1250

[7] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Ling.* 7 (2019), 597–610.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*. Retrieved from http://arxiv.org/abs/1409.0473.

[9] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, 178–186. Retrieved from http://www.aclweb.org/anthology/W13-2322.

[10] Tamali Banerjee, Anoop Kunchukuttan, and Pushpak Bhattacharya. 2018. Multilingual Indian language translation system at WAT 2018: Many-to-one phrase-based SMT. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation*. Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/Y18-3013.

[11] Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 1538–1548. DOI: https://doi.org/10.18653/v1/D19-1165

[12] Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup. 2016. Conditional computation in neural networks for faster models. In *Proceedings of the International Conference on Learning Representations (ICLR'16) Workshop Track*.

[13] Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 3112–3122. Retrieved from http://aclweb.org/anthology/C18-1263.

[14] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT'17). In *Proceedings of the 2nd Conference on Machine Translation*. Association for Computational Linguistics, 169–214. Retrieved from http://www.aclweb.org/anthology/W17-4717.

[15] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT'18). In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, 272–303. Retrieved from http://aclweb.org/anthology/W18-6401.

[16] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*. 2–14.

[17] Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 1853–1861.

[18] Rajen Chatterjee, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source neural automatic post-editing: FBK's participation in the WMT 2017 APE shared task. In *Proceedings of the 2nd Conference on Machine Translation*. Association for Computational Linguistics. 630–638. DOI: https://doi.org/10.18653/v1/W17-4773

[19] Aditi Chaudhary, Siddharth Dalmia, Junjie Hu, Xinjian Li, Austin Matthews, Aldrian Obaja Muis, Naoki Otani, Shruti Rijhwani, Zaid Sheikh, Nidhi Vyas, Xinyi Wang, Jiateng Xie, Ruochen Xu, Chunting Zhou, Peter J. Jansen, Yiming Yang, Lori Levin, Florian Metze, Teruko Mitamura, David R. Mortensen, Graham Neubig, Eduard Hovy, Alan W. Black, Jaime Carbonell, Graham V. Horwood, Shabnam Tafreshi, Mona Diab, Efsun S. Kayi, Noura Farra, and Kathleen McKeown. 2019. The ARIEL-CMU systems for LoReHLT18. *CoRR* abs/1902.08899 (2019).

[20] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3098–3112. DOI: https://doi.org/10.18653/v1/P19-1299

[21] Yun Chen, Yang Liu, Yong Cheng, and Victor O. K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1925–1935. DOI: https://doi.org/10.18653/v1/P17-1176

[22] Yun Chen, Yang Liu, and Victor O. K. Li. 2018. Zero-resource neural machine translation with multi-agent communication game. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI Press, 5086–5093.

[23] Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 3974–3980. DOI: https://doi.org/10.24963/ijcai.2017/555

[24] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST'14)*. Association for Computational Linguistics, 103–111. DOI: https://doi.org/10.3115/v1/W14-4012

[25] Gyu Hyeon Choi, Jong Hun Shin, and Young Kil Kim. 2018. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resource Association, 900–904. Retrieved from http://aclweb.org/anthology/L18-1144.

[26] Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Lang. Resour. Eval.* 49, 2 (2015), 375–395.

[27] Chenhui Chu and Raj Dabre. 2018. Multilingual and multi-domain adaptation for neural machine translation. In *Proceedings of the 24th Meeting of the Association for Natural Language Processing (NLP'18)*. 909–912.

[28] Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *CoRR* abs/1906.07978 (2019).

[29] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 385–391. DOI: https://doi.org/10.18653/v1/P17-2061

[30] Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 1304–1319. Retrieved from http://aclweb.org/anthology/C18-1111.

[31] Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, 531–540. DOI: https://doi.org/10.3115/1219840.1219906

[32] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 32nd Conference on Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 7059–7069. Retrieved from http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf.

[33] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://github.com/facebookresearch/MUSE.

[34] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2475–2485. Retrieved from https://www.aclweb.org/anthology/D18-1269.

[35] Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, 99–107. DOI: https://doi.org/10.18653/v1/D19-5610

[36] Raj Dabre, Fabien Cromieres, and Sadao Kurohashi. 2017. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. In *Proceedings of the Machine Translation Summit XVI, Vol.1: Research Track*. 96–106.

[37] Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 1410–1416. DOI: https://doi.org/10.18653/v1/D19-1146

[38] Raj Dabre, Anoop Kunchukuttan, Atsushi Fujita, and Eiichiro Sumita. 2018. NICT's participation in WAT 2018: Approaches using multilingualism and recurrently stacked layers. In *Proceedings of the 5th Workshop on Asian Language Translation*.

[39] Raj Dabre and Sadao Kurohashi. 2017. MMCR4NLP: Multilingual multiway corpora repository for natural language processing. *arXiv preprint arXiv:1710.01025* (2017).

[40] Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*. The National University (Philippines), 282–286. Retrieved from http://aclweb.org/anthology/Y17-1038.

[41] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: A multilingual speech translation corpus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics. 2012–2017. Retrieved from https://www.aclweb.org/anthology/N19-1202.

[42] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 1723–1732. DOI: https://doi.org/10.3115/v1/P15-1166

[43] Bonnie J. Dorr. 1987. UNITRAN: An interlingua approach to machine translation. In *Proceedings of the 6th Conference of the American Association of Artificial Intelligence*.

[44] Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 678–683. Retrieved from http://www.aclweb.org/anthology/P13-2119.

[45] Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. From bilingual to multilingual neural machine translation by incremental training. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*.

[46] Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of NMT-derived interlingual embeddings and their use in parallel sentence identification. *IEEE J. Select. Topics Sig. Proc.* 11, 8 (Dec. 2017), 1340–1350. DOI: https://doi.org/10.1109/JSTSP.2017.2764273

[47] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. 1126–1135. Retrieved from http://proceedings.mlr.press/v70/finn17a.html.

[48] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 866–875. DOI: https://doi.org/10.18653/v1/N16-1101

[49] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 268–277. DOI: https://doi.org/10.18653/v1/D16-1026

[50] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 1 (2016), 2096–2030.

[51] Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING'16)*. The COLING 2016 Organizing Committee, 1409–1418. Retrieved from http://aclweb.org/anthology/C16-1133.

[52] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. 1243–1252. Retrieved from http://proceedings.mlr.press/v70/gehring17a.html.

[53] Adrià De Gispert and José B. Mariño. 2006. Catalan-English statistical machine translation without parallel corpus: Bridging through Spanish. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. 65–68.

[54] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 344–354. DOI: https://doi.org/10.18653/v1/N18-1032

[55] Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3622–3631. Retrieved from http://aclweb.org/anthology/D18-1398.

[56] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1258–1268. DOI: https://doi.org/10.18653/v1/P19-1121

[57] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 6098–6111. DOI: https://doi.org/10.18653/v1/D19-1632

[58] Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation*. 1–7.

[59] Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2017. Effective strategies in zero-shot neural machine translation. In *Proceedings of the 14th International Workshop on Spoken Language Translation*. 105–112.

[60] Barry Haddow and Faheem Kirefu. 2020. PMIndia—A collection of parallel corpora of languages of India. *arxiv 2001.09907* (2020).

[61] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*. Retrieved from https://openreview.net/forum?id=SJgdnAVKDH.

[62] Carlos Henríquez, Marta R. Costa-jussá, Rafael E. Banchs, Lluis Formiga, and José B. Mariño. 2011. Pivot strategies as an alternative for statistical machine translation tasks involving Iberian languages. In *Proceedings of the Workshop on Iberian Cross-language Natural Language Processing Tasks (ICL'11)*. 22–27.

[63] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'14) Deep Learning Workshop*.

[64] Chris Hokamp, John Glover, and Demian Gholipour Ghalandari. 2019. Evaluating the supervised and zero-shot performance of multi-lingual translation models. In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, 209–217. DOI: https://doi.org/10.18653/v1/W19-5319

[65] Yanping Huang, Yonglong Cheng, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, and Zhifeng Chen. 2019. GPipe: Efficient training of giant neural networks using pipeline parallelism. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'19)*.

[66] Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: A geometric approach. *Trans. Assoc. Comput. Ling.* 7 (2019), 107–120. Retrieved from https://www.aclweb.org/anthology/Q19-1007.

[67] Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning. In *Proceedings of the Continual Learning Workshop at NeurIPS'18*.

[68] Girish Nath Jha. 2010. The TDIL program and the Indian language Corpora Intitiative (ILCI). In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/874_Paper.pdf.

[69] Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

[70] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Ling.* 5 (2017), 339–351. Retrieved from http://aclweb.org/anthology/Q17-1024.

[71] Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*.

[72] Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 866–876. DOI: https://doi.org/10.18653/v1/D19-1080

[73] Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1317–1327. DOI: https://doi.org/10.18653/v1/D16-1139

[74] Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Trans. Assoc. Comput. Ling.* 6 (2018), 225–240. DOI: https://doi.org/10.1162/tacl_a_00017

[75] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING'12)*. The COLING 2012 Organizing Committee, 1459–1474. Retrieved from https://www.aclweb.org/anthology/C12-1089.

[76] Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*. Association for Computational Linguistics, 244–252. Retrieved from http://www.aclweb.org/anthology/W18-6325.

[77] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*. AAMT, 79–86. Retrieved from http://mt-archive.info/MTS-2005-Koehn.pdf.

[78] Philipp Koehn. 2017. Neural machine translation. *CoRR* abs/1709.07809 (2017).

[79] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, 177–180. Retrieved from http://www.aclweb.org/anthology/P/P07/P07-2045.

[80] Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the 1st Workshop on Neural Machine Translation*. Association for Computational Linguistics, 28–39. Retrieved from http://www.aclweb.org/anthology/W17-3204.

[81] Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 127–133. Retrieved from https://www.aclweb.org/anthology/N03-1017.

[82] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 66–71. DOI: https://doi.org/10.18653/v1/D18-2012

[83] Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 1565–1575. DOI: https://doi.org/10.18653/v1/D19-1167

[84] Anoop Kunchukuttan. 2020. IndoWordnet Parallel Corpus. Retrieved from https://github.com/anoopkunchukuttan/indowordnet_parallel.

[85] Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Orthographic syllable as basic unit for SMT between related languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1912–1917. DOI: https://doi.org/10.18653/v1/D16-1196

[86] Anoop Kunchukuttan and Pushpak Bhattacharyya. 2017. Learning variable length units for SMT between related languages via byte pair encoding. In *Proceedings of the 1st Workshop on Subword and Character Level Models in NLP*. Association for Computational Linguistics, 14–24. DOI: https://doi.org/10.18653/v1/W17-4102

[87] Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-Anuvadak: Tackling multiway translation of Indian languages. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 1781–1787. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/414_Paper.pdf.

[88] Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, and Pushpak Bhattacharyya. 2017. Utilizing lexical similarity between related, low-resource languages for pivot-based SMT. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Asian Federation of Natural Language Processing, 283–289. Retrieved from http://aclweb.org/anthology/I17-2048.

[89] Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 641–652. Retrieved from http://aclweb.org/anthology/C18-1054.

[90] Surafel Melaku Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT'18)*. 54–61.

[91] Surafel Melaku Lakew, Quintino F. Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Improving zero-shot translation of low-resource languages. In *Proceedings of the 14th International Workshop on Spoken Language Translation*. 113–119.

[92] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rkYTTf-AZ.

[93] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Trans. Assoc. Comput. Ling.* 5 (2017), 365–378. Retrieved from http://aclweb.org/anthology/Q17-1026.

[94] Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, 84–92. Retrieved from http://aclweb.org/anthology/W18-6309.

[95] Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource NMT using multiple high-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Proc.* 18, 4 (May 2019). DOI: https://doi.org/10.1145/3314945

[96] Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2529–2535. DOI: https://doi.org/10.18653/v1/D17-1268

[97] Giulia Mattoni, Pat Nagle, Carlos Collantes, and Dimitar Shterionov. 2017. Zero-shot translation for Indian languages with sparse data. In *Proceedings of Machine Translation Summit XVI, Vol. 2: Users and Translators Track*. 1–10.

[98] Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. 33–40. Retrieved from https://www.aclweb.org/anthology/E06-1005.

[99] Cettolo Mauro, Girardi Christian, and Federico Marcello. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of European Association for Machine Translation*. 261–268.

[100] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR* abs/1309.4168 (2013).

[101] Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 3868–3873. Retrieved from https://www.aclweb.org/anthology/N19-1387.

[102] Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT'18)*. 1–41.

[103] Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1358–1367. Retrieved from https://www.aclweb.org/anthology/D09-1141.

[104] Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. *CoRR* abs/1703.01619 (2017).

[105] Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 875–880. Retrieved from http://aclweb.org/anthology/D18-1103.

[106] Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Asian Federation of Natural Language Processing, 296–301. Retrieved from http://aclweb.org/anthology/I17-2050.

[107] Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with missing data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics, 92–99. Retrieved from http://aclweb.org/anthology/W18-2711.

[108] Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with data augmentation. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT'18)*. 48–53. Retrieved from https://arxiv.org/abs/1810.06826.

[109] Eric Nyberg, Teruko Mitamura, and Jaime Carbonell. 1997. The KANT machine translation system: From R&D to initial deployment. In *Proceedings of the LISA Workshop on Integrating Advanced Translation Technology*. 1–7.

[110] Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of the Machine Translation Summit*, Vol. 8. 253–258.

[111] Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 644–649. Retrieved from https://www.aclweb.org/anthology/E17-2102.

[112] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (Oct. 2010), 1345–1359. DOI: https://doi.org/10.1109/TKDE.2009.191

[113] Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the 4th Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, 13–23. DOI: https://doi.org/10.18653/v1/W19-5202

[114] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4996–5001. DOI: https://doi.org/10.18653/v1/P19-1493

[115] Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 425–435. Retrieved from http://aclweb.org/anthology/D18-1039.

[116] Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 401–409.

[117] Raj Noel Dabre Prasanna. 2018. *Exploiting Multilingualism and Transfer Learning for Low Resource Machine Translation*. Ph.D. Dissertation. Kyoto University. Retrieved from http://hdl.handle.net/2433/232411.

[118] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 30th Conference on Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 6076–6085. Retrieved from http://papers.nips.cc/paper/7188-svcca-singular-vector-canonical-correlation-analysis-for-deep-learning-dynamics-and-interpretability.pdf.

[119] Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 383–391. DOI: https://doi.org/10.18653/v1/D17-1039

[120] Ananthakrishnan Ramanathan, Jayprasad Hegde, Ritesh Shah, Pushpak Bhattacharyya, and M. Sasikumar. 2008. Simple syntactic and morphological processing can help English-Hindi statistical machine translation. In *Proceedings of the International Joint Conference on Natural Language Processing*.

[121] Matīss Rikters, Mārcis Pinnis, and Rihards Krišlauks. 2018. Training and adapting multilingual NMT for less-resourced and morphologically rich languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA), 3766–3773.

[122] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *CoRR* abs/1609.04747 (2016).

[123] Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, 261–271. Retrieved from http://aclweb.org/anthology/W18-6327.

[124] Amrita Saha, Mitesh M. Khapra, Sarath Chandar, Janarthanan Rajendran, and Kyunghyun Cho. 2016. A correlational encoder decoder architecture for pivot based sequence generation. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING'16)*. The COLING 2016 Organizing Committee, 109–118. Retrieved from https://www.aclweb.org/anthology/C16-1011.

[125] Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31, 1 (1966), 1–10.

[126] Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL'09)*. Association for Computational Linguistics, 719–727. Retrieved from https://www.aclweb.org/anthology/E09-1082.

[127] Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12)*. IEEE, 5149–5152. Retrieved from http://dblp.uni-trier.de/db/conf/icassp/icassp2012.html#SchusterN12.

[128] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. *CoRR* abs/1907.05791 (2019).

[129] Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*.

[130] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 86–96. Retrieved from http://www.aclweb.org/anthology/P16-1009.

[131] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1715–1725. Retrieved from http://www.aclweb.org/anthology/P16-1162.

[132] Lierni Sestorain, Massimiliano Ciaramita, Christian Buck, and Thomas Hofmann. 2018. Zero-shot dual machine translation. *CoRR* abs/1805.10338 (2018).

[133] Petr Sgall and Jarmila Panevová. 1987. Machine translation, linguistics, and interlingua. In *Proceedings of the 3rd Conference on European Chapter of the Association for Computational Linguistics (EACL'87)*. Association for Computational Linguistics, 99–103. DOI: https://doi.org/10.3115/976858.976876

[134] Itamar Shatz. 2016. Native language influence during second language acquisition: A large-scale learner corpus analysis. In *Proceedings of the Pacific Second Language Research Forum (PacSLRF'16)*. 175–180.

[135] Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'20)*.

[136] Shashank Siripragrada, Jerin Philip, Vinay P. Namboodiri, and C. V. Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 3743–3751. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.462.

[137] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 778–788. DOI: https://doi.org/10.18653/v1/P18-1072

[138] Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 449–459. Retrieved from https://www.aclweb.org/anthology/N19-1044.

[139] Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the European Union's highly multilingual parallel corpora. *Lang. Resour. Eval.* 48, 4 (2014), 679–707.

[140] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*. The MIT Press, 3104–3112. Retrieved from http://dl.acm.org/citation.cfm?id=2969033.2969173.

[141] Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 963–973. DOI: https://doi.org/10.18653/v1/D19-1089

[142] Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*. Retrieved from http://arxiv.org/abs/1902.10461.

[143] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew M. Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 1574–1578.

[144] Jörg Tiedemann. 2012. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 141–151. Retrieved from https://www.aclweb.org/anthology/E12-1015.

[145] Jörg Tiedemann. 2012. Parallel data, tools, and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), 2214–2218. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

[146] Hiroshi Uchida. 1996. UNL: Universal networking language—An electronic language for communication, understanding, and collaboration. In *UNU/IAS/UNL Center*. Retrieved from https://www.semanticscholar.org/paper/UNL%3A-Universal-Networking-Language-An-Electronic-Uchida/f281c6a61ee69e4fa0f15f3f6d03faeee7a74e10.

[147] Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 484–491. Retrieved from https://www.aclweb.org/anthology/N07-1061.

[148] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 30th Conference on Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008. Retrieved from http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[149] Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2018. Multilingual NMT with a language-independent attention bridge. *CoRR* abs/1811.00498 (2018).

[150] David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the 2nd Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 33–39. Retrieved from https://www.aclweb.org/anthology/W07-0705.

[151] Karthik Visweswariah, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthakrishnan Ramanathan, and Jiri Navratil. 2011. A word reordering model for improved machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 486–496. Retrieved from https://www.aclweb.org/anthology/D11-1045.

[152] Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 560–566. Retrieved from http://aclweb.org/anthology/P17-2089.

[153] Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1482–1488. DOI: https://doi.org/10.18653/v1/D17-1155

[154] Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5823–5828. DOI: https://doi.org/10.18653/v1/P19-1583

[155] Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*. Retrieved from https://arxiv.org/abs/1902.03499.

[156] Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2955–2960. Retrieved from http://aclweb.org/anthology/D18-1326.

[157] Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1213–1223. DOI: https://doi.org/10.18653/v1/P19-1117

[158] Toon Witkam. 2006. History and heritage of the DLT (Distributed Language Translation) project. In *Utrecht, The Netherlands: Private Publication*. 1–11. Retrieved from http://www.mt-archive.info/Witkam-2006.pdf.

[159] Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Mach. Translat.* 21, 3 (2007), 165–181.

[160] Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 154–162. Retrieved from https://www.aclweb.org/anthology/P09-1018.

[161] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144 (2016).

[162] Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. COLING, 508–514. Retrieved from https://www.aclweb.org/anthology/C04-1073.

[163] Chang Xu, Tao Qin, Gang Wang, and Tie-Yan Liu. 2019. Polygon-Net: A general framework for jointly boosting multiple unsupervised neural machine translation models. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. International Joint Conferences on Artificial Intelligence Organization, 5320–5326. DOI: https://doi.org/10.24963/ijcai.2019/739

[164] Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 656–661. Retrieved from http://aclweb.org/anthology/P18-2104.

[165] Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. Exploiting pre-ordering for neural machine translation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L18-1143.

[166] Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 378–384. DOI: https://doi.org/10.18653/v1/P17-2060

[167] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 3530–3534. Retrieved from https://www.aclweb.org/anthology/L16-1561.

[168] Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 30–34. DOI: https://doi.org/10.18653/v1/N16-1004

[169] Barret Zoph and Quoc V. Le. 2017. Neural architecture search with reinforcement learning. In *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*. Retrieved from https://openreview.net/forum?id=r1Ue8Hcxg.

[170] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1568–1575. DOI: https://doi.org/10.18653/v1/D16-1163

Received July 2019; revised June 2020; accepted June 2020