School of Informatics

Informatics Research Review
Leveraging Auxiliary Languages for Low-resource Neural Machine Translation

B259531
January 2025

Abstract

Modern neural machine translation systems have achieved remarkable translation accuracy in the recent years. However, these neural networks often require large-scale datasets to train, which is an obstacle for low-resource languages which have very small amounts of data available. In this survey, we explore three methods which effectively employ auxiliary languages to assist in translating the resource-poor language pairs, namely multilingual training, transfer learning, and pivot translation. We discuss past and recent works on these methods and suggest future research directions to further improve low-resource translation.

Date: Thursday 23$^{\text{rd}}$ January, 2025

Supervisor: Bjoern Franke

1 Introduction

There have been significant advances in modern machine translation systems in the past decade, in some cases attaining human-level performance. However, these superior results only hold for about 20 common languages out of the 7000 languages in the world [1]. Commercial translation tools such as Google Translate and Microsoft Translator only support 100+ languages, leaving the vast majority of languages greatly overlooked. In this report, we present a survey of methods which exploit auxiliary languages in order to improve neural machine translation models for low-resource languages.

1.1 Background

Neural machine translation. Popularized by Bahdanau et al. (2014) [2], neural machine translation (NMT) has proven successful for translation tasks and become the standard both in industry and academia. Unlike rule-based and statistical machine translation, NMT uses deep learning neural networks to automatically learn the relationship between languages without the need for linguistic expertise or manually defining rules and features for the model. As illustrated in Figure 1, an NMT model typically consists of an encoder which converts the input text in the source language into a sequence of vectors and a decoder which then uses that vector representation to produce a translated output text in the target language. Most modern NMT systems use an attention mechanism [2] which enables the encoder and decoder to focus on the most informative parts of the source sentence and avoids mapping the variable-length source text into a fixed-length vector representation. However, a major drawback of NMT systems is that they require large-scale datasets to train, which poses a challenge for low-resource languages with small amounts of data available.

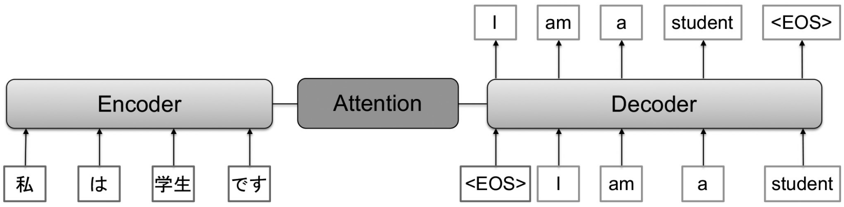

Figure 1: NMT encoder-decoder model architecture (original figure from [3])

Evaluation of NMT systems. Evaluating NMT models can be challenging, especially for languages with few native speakers. Manually assessing the quality of the translated texts is time-consuming and subjective. Moreover, human evaluation is difficult to scale as the number of languages and the size of the test set grow. Therefore, NMT research typically uses an automatic scoring method called “BiLingual Evaluation Understudy” (BLEU), introduced by Papineni et al. (2002) [4]. BLEU measures the closeness between the machine-translated text and a reference “gold-standard” translation. Most of the papers in this survey employed BLEU as the principal evaluation metric for NMT, while a few papers used both human judgment and automatic metrics.

Low-resource languages. There are several ways to define low-resource languages among researchers. In general, low-resource languages are understudied languages with scarce data resources, limited presence on the World Wide Web, and/or limited linguistic expertise [1]. In machine translation, the data scarcity issue is aggravated as parallel corpora from both the source and target languages are required to train the translation model. Recent literature tends to consider language pairs with less than 1 million parallel sentences as low-resource and less than 0.1 million parallel sentences as extremely low-resource [5]. For instance, even though there are sufficient monolingual datasets in English and in Balinese, translating between these two languages is treated a low-resource problem as there are only 0.3 million English - Balinese parallel sentences according to OPUS[1], the largest collection of public translation data.

Techniques for low-resource NMT. Numerous approaches have been proposed to improve the performance of NMT systems in a low-resource setting. These can broadly be grouped into four categories: data augmentation, modelling techniques, leveraging monolingual data, and leveraging data from auxiliary languages. This survey will focus on the use of auxiliary languages in low-resource NMT and will not cover the other categories of approaches. In practice, auxiliary languages are often high-resource languages with a level of relatedness to the source and/or target language of interest, which the NMT model can leverage to enrich the training data and enhance the translation quality. For instance, Zoph et al. (2016) [6] used French - English as a high-resource auxiliary language pair to assist in translation of the low-resource language pair Uzbek - English.

Leveraging auxiliary languages in low-resource NMT. Techniques which exploit auxiliary languages are further divided into three categories: multilingual training, transfer learning, and pivot translation [7]. Multilingual training builds a single model trained on parallel data from various languages and thus can translate between multiple language pairs. The intuition behind multilingual training is that the model can learn linguistic information from different languages which could be beneficial for learning the low-resource language pair of interest. Transfer learning involves training a "parent" model using an auxiliary language pair that is often high-resource, then transferring knowledge to a "child" model trained on the low-resource language pair. Pivot translation uses a "pivot" language as an intermediate translation step between the source and target languages. More specifically, the text is translated from source to pivot, then from pivot to target.

1.2 Overview and Structure

Our findings show that these methods are generally effective in leveraging information from auxiliary languages to improve machine translation for low-resource languages. However, in certain cases, the model capacity[2] and careful selection of auxiliary languages need to be taken into account. Overall, there is still a significant gap in the performance between low-resource and high-resource NMT due to the amount of data available, making low-resource translation an active area of research. Here, we present a comprehensive survey of state-of-the-art methods involving auxiliary languages in the context of NMT and suggest potential research directions to further improve low-resource translation.

This literature review is structured as follows: Section 2.1 will cover multilingual training for NMT in low-resource settings, Section 2.2 will discuss the research on transfer learning, and Section 2.3 will cover pivot translation. Finally, we will provide a summary in Section 3 and discuss current trends and future work in low-resource NMT.

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

[2] Model capacity roughly corresponds to the number of free parameters in the model.

2 Literature Review

2.1 Multilingual Training

The objective of multilingual training is to obtain a single model that is capable of translating multiple language pairs simultaneously. A major advantage of multilingual models is their ability to learn shared semantic representations across different languages, which allow low-resource languages to benefit from knowledge transfer from high-source languages [8]. Another advantage of multilingual models is their low complexity, as it is much simpler to train, maintain, and deploy one multilingual model than various bilingual models. Suppose we have sufficient training data for n languages, then we would need n^2 bilingual models to translate between all possible directions whereas with multilingual training, all translation directions can be performed by a single model.

The extent of parameter sharing in the encoder, decoder, and attention mechanism is an important decision in designing multilingual models. In terms of model architecture, multilingual models can take on one of the following forms, which will be discussed in this section:

- One-to-many: translating from one source language into multiple target languages
- Many-to-one: translating from multiple source languages into one target language
- Many-to-many: translating from multiple source languages into multiple target languages

2.1.1 One-to-many Multilingual Models

Dong et al. (2015) [9] first explored multilingual training in an one-to-many setting. In particular, they trained a unified model that can translate from English to Spanish, French, Portuguese and Dutch. This proposed model had one encoder and a separate decoder for each target language. For each of the four target languages, the multilingual model outperformed the respective bilingual model trained on a single language pair. Although the languages employed are high-resource, the authors simulated a low-resource setting by sampling 15% of the available data for each language pair. They found that the translation quality also improved in this low-resource setting and attributed this gain to the shared encoder's ability to leverage multiple resource-poor language pairs to learn a better representation of the source language.

2.1.2 Many-to-one Multilingual Models

Multi-source neural translation was introduced by Zoph and Knight (2016) [10], corresponding to a many-to-one NMT model. In this case, the source languages were French and German, while the target language was English. Each source language had its own encoder and attention mechanism, while the decoder was shared. The authors hypothesized that the French-to-English translation could benefit from the additional information in the German data, and likewise German-to-English translation could be boosted by the French knowledge. The multi-source models resulted in improvements over the single-source baselines. In addition, they were able to obtain greater gains when the source languages were more dissimilar as they provided distinct information from each other. However, this method was only tested in a high-resource setting.

Gu et al. (2018) [11] trained a many-to-one model that was especially effective in the low-resource setting. The proposed method involved sharing lexical information and sentence-level representations among the multiple source languages for translating into one target language. Since the input representations were shared among the source languages, knowledge from high-resource languages could be transferred to low-resource languages. This allowed some of source languages in the model to be low-resource or extremely low-resource, with very small amounts of parallel data with the target language. Moreover, using multiple source languages had a regularizing effect on the model and mitigated the issue of overfitting on a small language pair.

2.1.3 Many-to-many Multilingual Models

A many-to-many NMT model was studied by Firat et al. (2016) [12], whose training data contained parallel sentences from the WMT'15[3] dataset in English, French, Czech, German, Russian and Finnish. The model was considered a *multi-way* model as it could translate from any source language to any target language in the training corpus. There was a encoder for each source language and a decoder for each target language, while the attention mechanism was shared across languages. The motivation behind using a common attention mechanism was that knowledge could be transferred to from resource-rich to resource-poor languages. Based on the experiments, the multilingual model outperformed the single-pair models in both high-resource and low-resource settings, while requiring a considerably smaller number of model parameters compared to all of the single-pair models.

Johnson et al. (2017) [13] made simplifications to the multi-way model by using a single shared wordpiece vocabulary for all languages and shared model parameters for all language pairs. Hence, the model used the same encoder, decoder, and attention mechanism to represent the different languages. The motivation for sharing the parameters was to help the model learn a joint representation across all languages, which was beneficial for low-resource language pairs. The authors used an artificial token at the beginning of each source sentence to specify the target language. The proposed multilingual NMT model achieved comparable or slightly worse results than the individual single-pair models while requiring much less training time and fewer model parameters. A notable discovery was that the multilingual model demonstrated **zero-shot translation** capabilities, which enabled translation between language pairs that were not explicitly provided in the training dataset. In the paper, Portuguese was one of the source languages and Spanish was one of the target languages, but Portuguese - Spanish parallel sentences were not provided in the training data. The multilingual model was still able to translate from Portuguese to Spanish by leveraging its internal representations. The authors reported that the multilingual model by default had satisfactory zero-shot translation quality, but this could be improved by further training using a small number of true examples of the zero-shot language pair.

2.1.4 Massively Multilingual Models

Massively multilingual training seeks to push the scale of many-to-many NMT models in terms of number of languages without deterioration in translation quality. Aharoni et al. (2019) [14] trained massively multilingual models in both high-resource and low-resource settings. A limitation was that these proposed models were "English-centric" as they were trained on language pairs that had English as either the source or target language. In the high-resource setting, a single multilingual model was trained using 102 language pairs to-and-from English and up to 1 million training examples per language pair, corresponding to 204 translation directions. In the low-resource setting, the model was trained on 58 languages to-and-from English, resulting

[3] https://machinetranslate.org/wmt15

in 116 translation directions. During evaluation, the massively multilingual models outperformed the bilingual models of similar capacity in both low- and high-resource cases based on the BLEU scores. Training on various language pairs resulted in a regularizing effect which mitigated overfitting and improved model performance. The authors concluded that given sufficient data and model capacity, multilingual NMT models could effectively scale up to 103 high-resource languages and up to 59 low-resource languages in an English-centric setting while preserving high translation quality. In addition, the paper analyzed the trade-off between the number of languages and the translation quality. Given fixed model capacity, the translation quality generally decreased as the number of languages became very large. However, zero-shot translation improved as more languages were used for training the model.

Arivazhagan et al. (2019) [8] scaled the massively multilingual NMT models further by training a single model that could translate between 103 languages to-and-from English, using over 25 billion training examples. There was a substantial data imbalance as more languages were added to the system, since there are much fewer examples of low-resource language pairs than high-resource language pairs. The authors described a trade-off between positive knowledge *transfer* (improvement) to the low-resource languages and negative *interference* (degradation) to the high-resource languages as the number of languages increased. One of the factors that could impact the level of negative interference was the model capacity: as more languages were added to the model, the number of model parameters also needed to be increased. Similar to [14], zero-shot performance improved as more languages were involved in the model, especially among pairs of similar languages such as Belarusian and Russian.

Although massively multilingual models have attained respectable translation quality, the existing models were English-centric. To address this, Fan et al. (2021) [15] developed a non-English-centric multilingual model that could translate between 9,900 directions of 100 difference languages. To this end, a large-scale parallel corpus was mined from the Internet. When evaluated on the non-English directions, this proposed many-to-many model achieved a significant improvement in BLEU scores compared the English-centric baseline models, while still maintaining competitive performance on English-related translation directions. The authors highlighted using high model capacity as an important factor to effectively handle a very large dataset of many languages and translation directions.

Bapna et al. (2022) [16] tested the upper limits of multilingual training by building a massively multilingual NMT model that was capable of translating between 1000 languages in any direction. This was achieved by using a large-scale web-mined dataset containing clean, monolingual texts for over 1000 languages in addition to a parallel corpus of 112 languages. Similar to [15], the authors noted that it was crucial to have sufficiently high model capacity as the number of languages increased. Moreover, zero-shot translation improved in performance as more languages were added to the system, which was in agreement with [8]. Based on automatic metrics and human evaluation by native speakers on a subset of the languages, the model was able to attain satisfactory translation quality on long-tail languages which are extremely low-resource.

2.2 Transfer Learning

Transfer learning involves first training a parent model using an auxiliary language pair that is often high-resource. Next, knowledge from this parent model is transferred to a child model trained on a low-resource language pair of interest, as depicted in Figure 2. Knowledge transfer can be implemented by copying the word embeddings from the parent model to the child model or by *fine-tuning* which involves further training the parent model with the child language pair. Since the child model is able to learn from additional knowledge of the parent model, the child model often improves in translation quality while requiring much less parallel data for training.

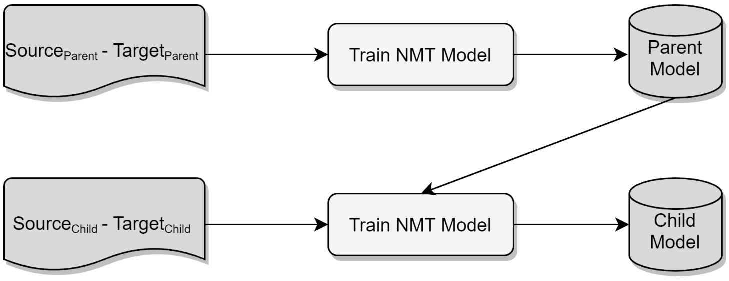
Source$_{Parent}$ - Target$_{Parent}$
Train NMT Model
Parent Model
Source$_{Child}$ - Target$_{Child}$
Train NMT Model
Child Model

Figure 2: Transfer learning procedure (original figure from [17])

Most of the literature suggests that transfer learning is most effective when the parent language pair is related to the child language pair [3]. This allows for greater vocabulary overlap between the parent and child models, which is advantageous for knowledge transfer. However, empirical results from [18] demonstrated that transfer learning is still useful when using an unrelated parent language pair, provided that there is a sufficiently large parallel corpus to train a strong parent model. In this section, we will discuss two types of transfer learning in low-resource NMT: lexical transfer and fine-tuning.

2.2.1 Transfer Learning via Lexical Transfer

Transfer learning was introduced by Zoph et al. (2016) [10] as an effective approach to exploit auxiliary languages. The method involved first training a parent model using a high-resource language pair, then using the some of the learned parameters of the trained parent model to initialize a child model which would then be trained on a low-resource language pair with the same target language. In the paper, the parent model was trained to translate from French to English, which is a resource-rich pair. There were four child models, respectively translating from Hausa, Turkish, Uzbek, Urdu to English. For the shared target language, the English word embeddings from the trained parent model were transferred to each child model and remained fixed. For the source languages, the word embeddings were initialized as random French word embeddings then iteratively updated as the child models were being trained. This approach allowed for knowledge transfer from a high-resource language pair to a low-resource language pair without re-training the parent model.

Nguyen and Chiang (2017) [19] extended transfer learning by making use of language relatedness between the parent and child language pairs. Unlike in Zoph et al. (2016) [10] which required a high-resource parent language pair, here the parent language pair could be another low-resource pair that had sufficient vocabulary overlap with the child language pair. This way, word embeddings of both the source and target languages could be directly copied from the parent model to the child model. Words were split using Byte Pair Encoding (BPE) [20] to further increase the overlap in lexicons between the parent and child languages pairs.

2.2.2 Transfer Learning via Fine-tuning

Kocmi and Bojar (2018) [18] proposed a “trivial” variant of transfer learning in NMT. Instead of copying the word embeddings from the parent model to the child model, the proposed method simply changed the parallel corpus during training without modifying any model parameters. More specifically, the model was first trained using the parent language pair for a number of iterations. Then, that same model continued to train using the low-resource child language pair until it reached convergence. Significant gains were observed on the model for the resource-poor child language pair, especially when the child training corpus was very small. The authors also reported that this transfer learning method was beneficial for both related and unrelated language pairs, which suggested that the size of the parent training corpus matters more than the relatedness between the parent and child language pairs.

Some works such as Neubig and Hu (2018) [21] combined multilingual training and transfer learning to exploit the benefits of both methods. In this paper, the authors sought to adapt a large pre-trained multilingual model to a low-resource language pair. In particularly, the authors wanted to translate from a specific low-resource language of interest into English. The method used a massively multilingual "seed model" as the starting point, then continued training jointly on the low-resource pair as well as a similar high-resource pair. Fine-tuning on related high-source data in addition to the low-resource data of interest resulted in *similar-language regularization*, which prevented overfitting on the low-resource data. The child training corpus was a simple concatenation of the two corpora, consisting of a small number of low-resource examples and a large number of similar high-resource examples. Alternatively, balanced sampling could be employed to maintain a 1-to-1, 1-to-2, or 1-to-4 ratio between the low-resource language and the similar high-resource language. The authors experimented with two adaptation settings: (1) *warm start*, in which the original multilingual model had access to the low-resource data of interest and (2) *cold start*, where examples of the low-resource language were not provided to the original pre-trained model but only in the fine-tuning step. In both warm-start and cold-start cases, leveraging the related high-resource language alongside the low-resource language improved the final translation quality.

2.3 Pivot Translation

Pivot translation is typically employed in cases where there are zero or very small amounts of parallel data for the language pair of interest. Instead of directly translating from the source language to the target language, a pivot language can be used as a bridge between them. The standard pivot translation method is performed by first translating from source to pivot, then from pivot to target. In low-resource and extremely low-resource settings, pivot translation tends to outperform zero-shot translation [3]. For both of these methods, translation quality can be improved by further training on a small amount of true or synthetic examples of the low-resource language pair of interest [17].

In practice, pivot translation requires training and deploying two or more separate models instead of one, which increases the complexity of the system. Another notable drawback of pivot translation is the error propagation problem [7]. If the intermediate translated text in the pivot language is inaccurate, then the final translated output in the target language will also have poor quality. Therefore, careful selection of the pivot language is crucial, yet largely understudied. In many cases, the pivot language is chosen such that both the source-pivot and pivot-target language pairs are resource-rich in order to build strong models for both translation steps. It is also possible to use several pivot languages as multiple intermediate steps between the source and target languages, but the error propagation would become more severe.

Researchers have addressed these issues by designing variants of pivot translation which leverage the pivot language to obtain a single source-target model. Chen et al. (2017) [22] proposed using the pivot language to generate pseudo-parallel data for the source-target language pairs. The authors then trained a single model using this synthetic pseudo-parallel dataset to directly translate from the source language to the target language to avoid propagating the error through the multiple steps.

Kim et al. (2019) [23] combined pivot translation and transfer learning in order to mitigate error propagation. First, the source-pivot and pivot-target models were trained separately as in standard pivot translation. Then, the child model was initialized using the encoder from the source-pivot model and the decoder of the pivot-target model. At inference time, this child model was capable of translating directly from source to target in one step without the need for a pivot language.

3 Summary & Conclusion

In this report, we provide an overview of past and current methods which utilize auxiliary languages to assist in machine translation for low-resource language pairs. In particular, we present three categories of methods: multilingual training, transfer learning, and pivot translation. All three methods showed success in improving the translation quality in low-resource and extremely low-resource settings. Moreover, some works achieved positive gains by using a combination of these methods to simultaneously exploit their advantages.

Transfer learning is an efficient way to transfer knowledge from one or more high-resource parent language pair(s) to a low-resource language pair of interest [6]. Moreover, it is straightforward to adapt to a new child language pair without having to re-train the parent model. Pivot translation is especially useful in cases where there are zero or very small amounts of source-target data but there is a sufficiently large data for a related language which can act as a bridge between the source and target languages.

Multilingual training is the most prominent and well-studied method out of the three. The majority of research in this subject has been moving towards massively multilingual many-to-many models [8, 14, 15, 16]. This approach results in a single model that is capable of translating between hundreds or thousands of translation directions. Furthermore, massively multilingual models exhibit strong zero-shot translation capabilities by implicitly bridging language pairs that are not explicitly provided in the training data. Most research emphasized the importance of having high model capacity to effectively learn meaningful representations from large-scale multilingual datasets. Many papers reported the transfer-interference trade-off in multilingual models: as the number of languages grows, the translation performance improves for low-resource language pairs but decreases for high-source ones.

Some of the latest research efforts in low-resource NMT focused on non-English-centric datasets and models, which enable significantly more translation directions and use cases. Recent works have trained massively multilingual models capable of translating in any direction across 100 and 1000 languages [15, 16]. Goyal et al. (2022) [5] presented FLORES-101, a dataset of 3001 sentences which were translated into 101 languages by professional experts, resulting in 10,100 translation directions. The dataset was later expanded to FLORES-200 by the NLLB Team (2022) [24] to support manually translated sentences in 200 different languages which correspond to over 40,000 translation directions.

In the recent years, there has been a shift towards using high-quality, manually-validated data for training and evaluation in NMT systems. Parallel sentences in FLORES-101 [5] and FLORES-200 [24] were produced by professional translators to ensure very high data quality. Human evaluation was used as the final benchmark for translation performance in Bapna et al. (2022) [16], while Maillard et al. (2023) [25] showed that using a small training corpus of professionally translated sentence pairs was much more effective than using a very large corpus of lower quality.

The choice of auxiliary languages plays a key role in transfer learning and pivot translation, yet there is limited research on this topic. Future work could investigate how the level of relatedness as well as the number of auxiliary languages impact these NMT methods. Recent multilingual training efforts focused on building massive models capable of translating between tens of thousands of directions. As next steps, we could further study the transfer-interference trade-off [8] to better understand and mitigate the degradation in high-resource languages in multilingual models while still improving performance for low-resource languages. Furthermore, using monolingual data in conjunction with parallel data has shown good potential in multilingual models [16]. We could explore to what extent the monolingual data influences the model in comparison to the parallel corpus. Given the current trends in NMT, future studies could assess the impact of using small but curated, high-quality datasets compared to using very large but synthetic, lower-quality datasets.

References

[1] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

[2] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38, 2020.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[5] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

[6] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.

[7] Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. A survey on low-resource neural machine translation. *arXiv preprint arXiv:2107.04239*, 2021.

[8] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

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

[10] Barret Zoph and Kevin Knight. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*, 2016.

[11] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*, 2018.

[12] Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio. Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252, 2017.

[13] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[14] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*, 2019.

[15] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.

[16] Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*, 2022.

[17] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

[18] Tom Kocmi and Ondřej Bojar. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*, 2018.

[19] Toan Q Nguyen and David Chiang. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*, 2017.

[20] Rico Sennrich. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[21] Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*, 2018.

[22] Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. A teacher-student framework for zero-resource neural machine translation. *arXiv preprint arXiv:1705.00753*, 2017.

[23] Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*, 2019.

[24] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[25] Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, 2023.