IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 October 2023, accepted 20 November 2023, date of publication 23 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336019

SURVEY

Low-Resource Neural Machine Translation: A Systematic Literature Review

BİLGE KAĞAN YAZAR, DURMUŞ ÖZKAN ŞAHİN, AND ERDAL KILIÇ
Faculty of Engineering, Ondokuz Mayıs University, 55139 Samsun, Turkey
Corresponding author: Bilge Kağan Yazar (bilgekaganyazar@gmail.com)

ABSTRACT In this study, a systematic literature review was conducted to examine the significant works in the literature on low-resource neural machine translation. Within the scope of the study, three research questions were identified to examine the low-resource neural machine translation literature. According to the inclusion and exclusion criteria, 45 studies were selected for review. After the relevant studies were identified, three research questions were aimed to be answered. The first research question is to identify the study directions and language pairs used in low-resource neural machine translation. The second research question aims to identify which deep learning methods are used in low-resource neural machine translation and which metrics are used to evaluate these methods. The third research question is to determine the bilingual and monolingual corpora used in the studies and the preferred development environments. In addition, the studies with the most commonly used language pairs were analyzed, and directions for future studies were made.

INDEX TERMS Neural machine translation, low resource languages, evaluation criteria, deep learning.

I. INTRODUCTION

Machine translation (MT) is a concept proposed in 1949 by Warren Weaver, who thought that computers could be used to automatically translate one language into other languages [1]. MT is a field of study that has received great attention in recent years, as it has similar goals with natural language processing (NLP) and machine learning (ML) concepts. Apart from its scientific importance, MT also has great potential in the field of communication [2]. Before deep learning approaches were applied to the field of MT, generally rule-based and statistical machine translation methods were used.

Rule-based machine translation was the first MT concept based on the assumption that there are words in all languages that has the same meaning, and was a popular method before the 2000s [3]. In this method, translation can be considered as placing the words in the source sentence in the appropriate place in the target language. Since the meaning of a sentence may be represented by different word orders in different languages, such a word substitution method must comply with the syntax rules of the languages to be translated. In such methods, certain rules must be designed for source language analysis, translation from source language to target language and target sentence generation. However, since there are so many syntactic rules in a language, editing grammar rules in this way is a very difficult process and requires a lot of effort. Although rule-based methods look good in theory, they lag far behind in terms of performance in practice because the defined rules do not include invisible rules in the language. The main disadvantage of rule-based methods is that they ignore the need for contextual information in the translation process, which makes machine translation unreliable.

Statistical machine translation methods, proposed in the 1990s, are systems that can learn translation rules between words or phrases using probabilistic models [4]. It is a method that has achieved success in the sector, especially in large companies such as Google and Microsoft. Unlike the rule-based approach, SMT models consider the translation process from a statistical perspective. SMT models find words or phrases with the same meaning through bilingual parallel corpora. The most widely used form of SMT is phrase-based SMT [5], which roughly includes preprocessing, word alignment, sentence alignment, and language model (LM) training. The basis of this model is the use of a vocabulary that matches phrases between the source and the target

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

language. In this method, unlike rule-based methods, the translation model can use contextual information in the sentence. Although SMT gives better results than rule-based methods, the systems that need to be designed manually, such as the language model and reordering model, cause SMT to not take full advantage of the parallel corpora, and the translation quality is far from desired performance [6].

Traditional machine learning techniques rely primarily on human-generated features derived from linguistic intuition, which is a trial-and-error process and frequently far less accurate at capturing the core of the original data. SMT techniques have done pretty well in the MT area in recent years; however, certain fundamental shortcomings still need to be resolved. The first one is that since SMT methods create the translation by splitting the source sentence into several phrases and changing the phrases, they ignore the long-term dependencies in long sentences, therefore it causes inconsistencies in the translation results. Second, existing systems often have many complex sub-components, such as language model, reordering model, etc. It gets increasingly challenging to adjust and combine these sub-components to produce a more stable output as their number rises. These circumstances have caused an obstacle in the advancement of SMT architectures. This problem is mainly due to the LM component. LM is able to provide important information, such as the probability of a specific word (or phrase) occurring based on prior words. Therefore, creating a effective LM greatly affects translation performance.

While the research of LM components through statistical methods has become almost static, neural language models (NLM) using a neural network to model text data directly have emerged. Due to the distributed representation of the words, NLM reduces sample sparsity in comparison to classic LM, enabling them to share statistical weights rather than being independent variables. However, LMs created using feed-forward networks have some problems due to neural networks. The most important one is the long-term dependency problem in the sentences. Language models using recurrent neural networks (RNN) structures have been put forth as a solution to this issue [7]. This method processes each word in a one time step, and the whole sentence is modeled. Thus, real conditional probability can be modeled without the limit of the content window [8]. With language models built using RNN, any size input can be processed, and information from previous steps can be used, but the computations are slow as a result of the numerous parameters.

The use of neural networks for MT (Neural Machine Translation: NMT) operations has required many years due to the low performance of models and hardware limitations for the calculations. First studies were done to build NLMs for the target language [9] and to apply statistical models [10]. These ideas have been taken further, including systems that score sentence pairs with a forward network [11] and work that adds a source content window to neural language models [12], [13]. The use of deep learning approaches for MT has started with studies in the last 10 - 12 years. With the spread of deep learning in 2010, the field of NLP has shown great progress. However, the use of deep neural networks for MT has also become widespread. Deep learning-based approaches, which are a completely new approach to MT, were first introduced in 2013 [14], [15]. Compared to other models, NMT models require less grammar and produce at least as good results as other methods [16]. Numerous studies have shown that NMT outperforms traditional SMT models and is industrially applicable to a greater extent [17].

With the increasing success of deep learning in the field of NLP, nowadays NMT models are designed as end-to-end learning. That is, a sequence of words in the source language is directly mapped to a sequence of words in the target language. The purpose of the learning process is to obtain the target sentence by viewing the two sentences as a high-dimensional classification problem in a semantic space. Encoding and decoding are the two components that make up this process in contemporary NMT models.

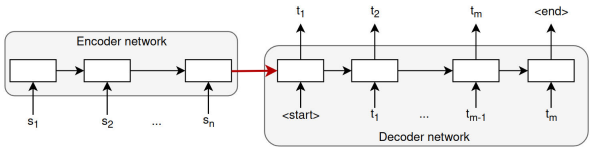


FIGURE 1. Example of basic encoder-decoder structure. The vector shown in red represents the encoding of the source sentence into a fixed-size vector.

An example visualization of the basic encoder-decoder structure is given in Figure 1. The encoder - decoder models generate the target $T = (t_1, t_2, \cdots, t_m)$ sentence using the maximum valued conditional probabilities in the source sentence $S = (s_1, s_2, \cdots, s_n)$. In doing so, it uses both predicted words and information from the source sentence. So this is an recurrent neural language model (RNLM) creation process. The encoder network sequentially processes a source sentence word by word upon receiving it, compressing the variable length sequence into a fixed length vector. The target sentence is subsequently generated by the decoder using the encoder's final hidden state. It is referred to as end-to-end translation because the encoder-decoder structure conducts translation directly from the source data to the target result, i.e. there is no obvious outcome in the intermediate step. The idea behind the encoder-decoder structure is to map the source sentence to the target sentence using a semantic space intermediate vector. The semantic meaning of both languages can be represented by this intermediate vector. RNN-based NMT models differ from one other in three key ways: (a) the way the sentence is given to the model; (b) the type of neural network used (SimpleRNN, LSTM, GRU); and (c) the depth of the RNN layer [18], [19]. Some models use CNN structure instead of RNN units in the encoder-decoder framework [20], [21]. There are various benefits to using convolution in NMT models rather than recurrence. Their hierarchical structure connects far-off words in the sentence more quickly than sequential

structures, requires fewer sequential calculations, and is easier to parallelize [22]. Because of these advantages, CNN-based models can facilitate the learning process. However, the models become deeper and more challenging to train when numerous convolution layers are stacked for translating long sentences [22].

The biggest problem in encoder-decoder structures is the process of compressing all the information in the source sentence into a fixed-size vector. This situation causes the performance of the models to decrease as the length of the sentence to be translated increases. In order to solve this problem, models with attention mechanisms that perform alignment and translation at the same time have been proposed. The first example of the attention mechanism was proposed in 2014 and has come to the fore as a very important development in the field of NLP [23]. While using an attention mechanism to generate each word in a translation model, it looks for a few places in the original sentence where the key information is concentrated. After that, using the content vectors connected to these source sentence positions and the words predicted in earlier time steps, the model predicts a new word. The most important feature of this structure; the source sentence does not need to be encoded into a fixed-size vector. Instead, the input is encoded as a sequence of vectors and a sub-set of these vectors is used in the decoding step. In this way, translation performance increases in long sentences. This method is used by the decoder to determine which elements of the source sentence should be given significance.

The attention mechanism has undergone many changes since the day it was first proposed and has been used in different ways. The attention mechanism in NMT is most frequently used as an interface between the encoder and the decoder, though. A significant refinement of the attention mechanism is the self-attention mechanism proposed in 2017 [24]. In the proposed structure named Transformer, RNN units have been removed and a structure that uses full attention has been created. Self-attention calculates the word dependency within the sentence sequence and thus obtains a stronger attention-based sequence representation. In the computational steps, self-attention first takes three vectors based on the original embedding for different purposes. The Query (Q), Key (K), and Value (V) vectors are the three vectors in question. Self-attention, which can be thought of as a mapping between Q, K, and V to output, is the Transformer's central element. Scaled-dot product attention and multi-head attention, two crucial attention mechanisms, are used to achieve this in the original Transformer study. These two key components of the Transformer model are depicted in Figure 2.

The dot-product of Q-queries and K-keys (size d_k) is calculated in the scaled-dot product attention process, and the outcome is scaled by divided $\sqrt{d_k}$. The results of the preceding phase are then put through the softmax function to produce the weights that will be multiplied by V. The attention output is calculated by multiplying these weights

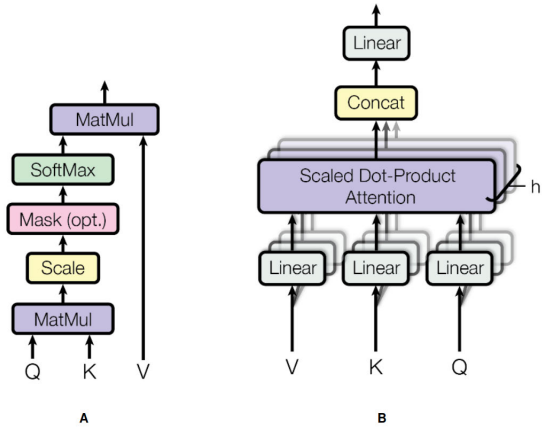


FIGURE 2. Attention mechanisms used in the Transformer model. A: Scaled dot-product attention, B: Multi-head attention [24].

by V. In practice, the attention computation is carried out concurrently over a series of queries in a Q matrix [24]. The matrices K and V are utilized to use keys and values. The formulation of this method is as follows [24]:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

The main idea in the Transformer structure is to perform as many operations as the number of attention heads (H, H = 8 in the original Transformer structure) instead of performing a single operation on the sentence. For attention heads, the query, key, and value vectors are linear transformations of Q, K, and V. The attention output is produced on each head using scaled dot-product attention. The combination of the outputs from each self-attention head is the result of multi-head attention. This is formulated as follows [24]:

$$MultiHead(Q, K, V) = Concat(head_1, \cdots, head_h)W^O \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

Here, all of the W matrices are parameters that can be learned. In this way, a much stronger representation is created and operations can be performed in parallel. The dimensions of the attention heads are usually divided by H to avoid increasing the number of parameters. Multiple sub-nets with diverse views of the key-value set running in parallel as multi-head attention sub-nets that process the output representation into various sub-spaces.

The Transformer model is displayed in Figure 3. Like earlier NMT models that have been successful in the literature, the Transformer model is built on the encoder-decoder structure. One of the difficulties encountered in self-attention-based models is that attention itself does not have a concept of order [22]. Key-value pairs are accessed only based on the correspondence between the key and the query, not based on the location of the key in memory. Since queries, keys, and values in recurrent NMT are obtained from RNN states and the RNN structure provides a strong sequential

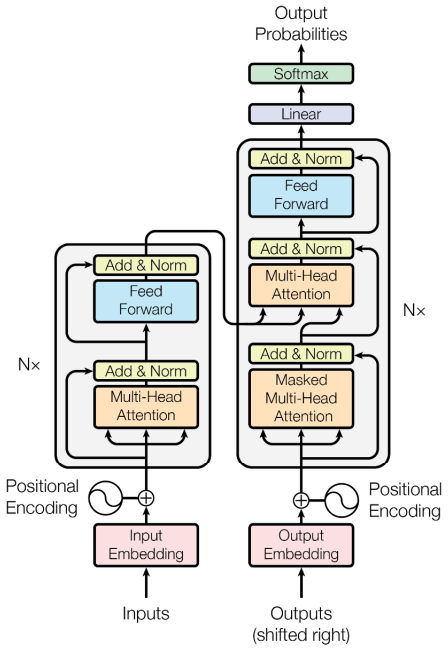


FIGURE 3. Transformer model [24].

signal, this does not present a significant challenge [22], [24]. Transformer model does not use recurrence; hence, handling the order of the words in the input sequences requires knowledge of the relative or absolute position of the tokens in the sequence. To overcome this, a method called positional embedding (PE) using sine and cosine functions is applied after the input and output embedding layers. By including them in the input and output word embeddings, these are become position-aware. This process is carried out as follows [24]:

$$PE_{pos,2i} = sin(pos/1000^{2i/d_{model}}) \quad (4)$$

$$PE_{pos,2i+1} = cos(pos/1000^{2i/d_{model}}) \quad (5)$$

After PE, The resulting output is then sent to the encoder. The Transformer encoder is a stack of N = 6 identical layers. Two sublayers make up each layer. A multi-head self-attention layer is the first sub-layer, while a fully connected feed-forward layer makes up the second. Each of these layers has a residual connection surrounding it, which is followed by a layer normalization operation. The Transformer decoder is a stack of N = 6 identical layers. The decoder features a third sub-layer that performs multi-head attention on the output of the encoder stack in addition to the two sub-layers in each encoder layer. The outputs are produced using residual connections and layer normalization, just as the encoder. In addition, the multi-head attention sub-layer is used in this part as masked multi-head attention. To stop the model from focusing on later tokens, subsequent embeddings are masked in this section. This ensures that at location t it can only use information from outputs generated from locations before t. Once the output is obtained from the decoder layer, it moves to the inference stage, where a softmax layer is used to generate the target sentence.

Since 2013, neural networks using the encoder - decoder system have become mainstream for MT studies. Today, it stands out as the technique used in Transformer architecture and the most used technique for NMT studies. Unlike other NLP methods, MT includes two languages. Therefore, the success of the model created in MT on a language pair is highly dependent on the number of parallel sentences available between the two languages. In order for NMT systems to achieve smooth results, large amounts of parallel data are needed in the created systems. High-resource language pairs (English, German, French, etc.) have no problem with parallel data. However, this is not the case for low-resource languages, and this is considered a major challenge for the NMT field. As a result, NMT research on low-resource languages has significantly increased in recent years. In NLP, the problem of low-resource is mainly due to low-resource of the considered languages or low-resource of the studied areas [25], [26]. Whether a language is low-resource or high-resource can be determined based on the size of data available and the NLP tools that can be used [25], [26], [27]. Additionally, a language is regarded as low-resource for NMT even if it involves a large number of monolingual corpora and a little parallel corpus with another language.

The main purpose of this study is to perform a systematic literature review (SLR) on NLP and deep learning methods used in low-resource NMT. Although there are many research studies examining on these topics on low-resource NMT, there are very few systematic reviews on this subject as far as it is known. Research articles for our study were carefully selected to examine the deep learning techniques used for low-resource NMT and the NLP methods used.

The remainder of this study consists of five parts: In Part II, a literature review is given. The methodology for how the studies reviewed in this study were obtained are described in Chapter III. Findings and evaluations are shared in Chapter IV. Chapter V includes discussion and conclusion.

II. LITERATURE REVIEW

When the studies in the field of low-resource NMT are analyzed, it is seen that the methods used utilize monolingual and auxiliary language data in addition to the limited corpus available. This section will examine the most widely used methods in low-resource NMT in general terms.

A. USE OF MONOLINGUAL DATA

Low-resource language pairs often perform poorly on the MT task due to the lack of parallel bilingual data. To address this issue, the use of monolingual data is recognized as an effective strategy to improve translation quality in low-resource scenarios. Monolingual data is especially helpful for enhancing translation accuracy in low-resource scenarios since it is more abundant and simpler to get than bilingual parallel data and provides a wealth of linguistic and contextual

information. Many studies have made extensive use of monolingual data in NMT systems, which are categorized in several aspects.

One of the most used method of monolingual data is back-translation (BT). Back-translation is the reverse translation of monolingual sentences from the target side into the source language using a translation system to create pseudo-parallel sentence pairs [28], [29]. When using this method, it has been shown that translating target sentences into source sentences usually yields better results [26]. An essential limitation of the BT method is the assumption that there is an NMT system in the BT direction, and the success of the NMT system used affects the model to be created. In addition, the synthetic data generated using BT contains more noise than the original data. Following this method, an iterative BT method has been proposed in the literature, which is based on BT and improves the success of NMT [30]. In the iterative method, the source and target data are translated using NMT models in opposite directions. This translation process is continued until there is no improvement on either side. There are many different back translation methods in the literature, and studies have shown that this method provides performance gains in NMT systems [29], [31].

Utilizing monolingual data and pre-trained models is helpful for a variety of language generation and understanding tasks [32], [33]. Since NMT requires both language understanding (encoder) and generation (decoder) capability, pre-training can be extremely beneficial, especially low-resource scenarios [34]. Depending on the encoder and decoder in the NMT, studies on language model pre-training can be categorized as separate or joint pre-training. Some studies use separate pre-training of the encoder and decoder. In [35], they experimented with initializing the encoder and decoder with different models, including BERT, GPT-2, RoBERTa, and random initialization. An LM can be added into the target side of the NMT model to increase the output text's fluency. This process is known as LM fusion, and classified into shallow fusion and deep fusion [32], [36]. In shallow fusion, LM is used to score words produced by the NMT system's decoder at inference time or during training [36]. The NMT design is changed in deep fusion, which improves performance, to integrate the LM and the decoder [36]. One drawback of the encoder and decoder used with separate pre-training is that they do not train the NMT well, which is crucial for linking source and target representations in the NMT model. To improve translation accuracy, some research suggest pre-training the encoder, decoder, and attention jointly [37], [38].

Recently, models utilizing adversarial training frameworks of unsupervised Generative Adversarial Network (GAN) structures with monolingual corpora and cross-lingual embeddings have become popular. For this structure, usually in the adversarial framework, initial translation models are created for both forward and backward directions, and then iterative BT is performed to improve translation performances jointly [26]. The neural network can learn a reliable map of the translation due to the adversarial training. The translation task is therefore framed by a generator and a discriminator using in a GAN architecture. Reconstruction loss is caused by reconstruction of forward and backward noisy translations [26]. The discrimination loss is a result of a binary classifier that distinguishes between the translated and original target texts in order to distinguish between the source language and the target language [26]. An adversarial loss function exchanges between the reconstruction loss of the back translation and the discrimination loss of the classifier. This process produces a superior translation that is more smooth for LRLs. Existing approaches in the literature on unsupervised NMT change the adversarial framework by incorporating extra adversarial phases or extra loss functions during the optimization step [39].

B. USING DATA IN THE AUXILIARY LANGUAGE

Human languages share similarities in several ways: languages in the same/similar language family or of a similar type can share similar writing style, vocabulary, and grammar; languages can affect one another, and a word from one language may be adopted as is in another [34]. In addition, translating between a low-resource language pair can be aided by a corpus of related languages [47]. The methods of utilizing data from different languages in low-resource NMT can be categorized as multi-lingual translation, transfer learning, and pivot translation.

The significant advantage of multi-lingual training is that multiple language pairs can be trained in a single model through parameter sharing. Compared to training multiple separate models, the cost of maintaining and model training can be significantly decreased, and information can be learned collectively from multiple languages to help LRLs [34], [58]. Low-resource language pairs can benefit from high-resource language pairs through joint training. When the languages in the models are linked, and the number of languages is relatively small, better results can be obtained than with bilingual models [26], [59]. Multi-lingual methods are more practical than building bilingual models because they include many languages. A review of the literature shows that multi-lingual methods can be modeled as one-to-many from one source language to many target languages, many-to-one from many source languages to one target language, or many-to-many from many source languages to many target languages [26], [43]. These methods are built by applying a single encoder-decoder, multiple encoders-single decoder, single encoder-multiple decoders, or multiple encoder-decoder models. Finally, multi-lingual NMT allows translation over language pairs not seen during training, so-called zero-shot translation (ZST) [60].

Transfer learning (TL) can be defined as the application of knowledge gained from solving one problem in machine learning to a different problem related to that problem. One of the most popular methods for low-resource NMT is TL. Initially, a NMT model is trained as a ''parent'' on language

TABLE 1. Comparison of other survey/review studies (Y: Yes, N: No, P: Partially).

Survey/Review	Year	Summary	Limitations	Systematic (Y/N/P)	Low-resource specific (Y/N/P)	Are metrics mentioned? (Y/N/P)	Are corpora mentioned? (Y/N/P)
Six Challenges for Neural Machine Translation [40]	2017	The 6 key challenges in NMT are examined by the authors. These are beam search problems, amount of training data, long sentences rare words, word alignment, and domain mismatch.	Other challenges in NMT are not addressed. In addition, limited studies on NMT were reviewed.	N	N	N	N
Machine Translation using Semantic Web Technologies : A Survey [41]	2018	The findings of a thorough analysis of Semantic Web-based machine translation methods for translating texts are presented.	The results of the reviewed articles are detailed not compared.	Y	N	Y	N
Neural machine translation: A review of methods, resources, and tools [42]	2020	General evaluations are made on NMT techniques. Architectures, decoding techniques and data augmentation approaches for NMT are described. In addition, the tools that are frequently used in this field are mentioned.	The results and comparisons of the studies examined are not included.	N	N	N	N
A Survey of Multi-lingual Neural Machine Translation [43]	2020	It is a compilation study on Multi-lingual Neural Machine Translation (MNMT). The MNMT area was evaluated under 3 groups. The techniques used in these structures are highlighted.	The results obtained and the metrics used are not detailed.	N	N	N	Y
Arabic Machine Translation: A survey of the latest trends and challenges [44]	2020	Translation studies of the Arabic language are discussed. The techniques used in the translation systems of the Arabic language and the results of the researches are included.	Languages other than Arabic were not studied.	N	N	N	Y
Neural machine translation: Challenges, progress and future [6]	2020	A review of the NMT framework is in progress. The challenges in NMT are also discussed. Recent developments in this field are presented. Finally, some potential future research trends are explored.	Evaluation is limited for corpora, tools, and metrics.	N	N	N	N
Machine Translation Systems for Indian Languages: Review of Modelling Techniques, Challenges, Open Issues and Future Research Directions [45]	2020	Machine translation systems on Indian languages are examined in this review. The models, difficulties and open problems used in the translation of Indian language pairs are examined.	It covers only machine translation techniques on Indian languages.	N	P	N	N
Recent Advances in Dialogue Machine Translation [46]	2021	It is a survey study on dialogue machine translation. Studies, current developments and corpora related to dialog machine translation are explained.	The results of the studies reviewed are limited. Results can be detailed in comparison.	N	N	N	Y
Recent advances of low-resource neural machine translation [47]	2021	It contains information on the latest developments in NMT research for low-resource languages. NMT studies were examined specifically for LRL. Some corpora and important conferences in this field are included in the compilation.	The results of the studies and the achievements are given in the article in a limited way.	N	Y	N	Y
A Survey on Document level Neural Machine Translation Methods and Evaluation [48]	2021	By giving the infrastructure of neural machine translation at the document level, important studies in this field are summarized.	Comparisons of the studies are given without detailing.	N	N	P	P
A Review of Machine Translation for South Asian Low Resource Languages [49]	2021	Machine translation approaches used for low resource languages are analyzed.	Few studies have been reviewed.	N	Y	N	N
Neural machine translation: past, present, and future [50]	2021	The development of NMT is described. In addition, the details and purposes of different attention mechanisms are given. The tools specific to this field used in the development of NMT systems are presented in the article. Finally, the advantages of NMT over SMT and current problems of NMT are discussed and the article is concluded.	The number of studies reviewed is limited. In addition, no detailed evaluation was made regarding the results of the studies examined.	N	N	P	P
Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey [51]	2022	Approaches to domain adaptation for NMT were examined.	Review on other NMT studies is limited.	N	N	N	N
Progress in Machine Translation [52]	2022	Progress in machine translation is included. Emphasis is placed on the development from early techniques to cutting-edge techniques in the field of machine translation.	Details of the studies are not included.	N	N	P	N
Survey of Low-Resource Machine Translation [53]	2022	Current studies on low-resource language machine translation studies are reviewed in this review.	The results of the studies are not detailed. Future directions are not made.	N	Y	N	Y
Neural Machine Translation for Low-Resource Languages: A Survey [26]	2023	Studies specific to low-resource NMT have been examined. The researchers working in this field were given guidance on the infrastructure of NMT and future studies.	The corpora used in the studies and the obtained performances are given as limited.	Y	Y	N	N
A Survey on Non-Autoregressive Generation for Neural Machine Translation and Beyond [54], A Survey of Non-Autoregressive Neural Machine Translation [55]	2023	Studies with non-autoregressive approach have been examined in detail.	There are no studies on the auto-regressive approach.	N	N	N	N
Machine translation status of Indian scheduled languages : A survey [56]	2023	The use of different languages spoken in India on machine translation systems has been examined.	Details such as the corpora used and the achievements are not detailed.	N	P	N	N
Transformer: A General Framework from Machine Translation to Others [57]	2023	This review study looks at the applications of Transformer-based NMT and Transformer structure in various areas as NLP, computer vision, and sound processing.	The results and comparisons of the studies examined are not included.	N	N	N	N
Our work	**2023**	**A systematic literature review was carried out for low-resource NMT. Language pairs used in studies in this area are included. In addition, the results of each translations are given on the BLEU score. Finally, the corpora and NMT tools used in the studies were presented to the researchers.**	**As a result of SLR, studies specific to low-resource NMT in Q1 and Q2 quarters (WOS) were examined.**	**Y**	**Y**	**Y**	**Y**

pairs that are resource-rich, typically with ample training data. Subsequently, this parent model is fine-tuned on a low-resource language pair, referred to as the ''child'' model, where training data is limited [61]. Fine-tuning is required to transfer information from the parent model to the child model. There are different approaches to fine-tuning, although it is unclear which method is better. These methods are; i) completely transferring the parent to the child, ii) fine-tuning the entire child model, iii) fine-tuning specific layers on the encoder-decoder models. The simplest way to fine-tune is to establish the model with a high-resource language pair, and then adjust the parameters using the low-resource language pair [62]. During fine-tuning, certain parameters can be fixed. This choice is purely a matter of model design. Furthermore, besides the bilingual parent model, using a multilingual parent model is another option [43]. Due to the restricted model capacity of a multilingual model, fine-tuning can drive the model to focus on the desired low-resource languages, boosting performance. As a result, a low-resource language pair can benefit from several auxiliary languages.

Typically, a high-resource language is selected as a bridge in pivot-based techniques. The source-target translation can then be constructed using the source-pivot, pivot-target corpus, and model. Often, source-pivot, pivot-target models are trained and then combined into a source-pivot-target model. [63]. Training the source-target model using pseudo-parallel data generated with the pivot language is another frequently used technique. Also, utilizing the parameters of the source-pivot and pivot-target models is one way of using the pivot language [64]. In pivot translation, the pivot language selection has a substantial impact on the translation's quality. A pivot language is typically selected based on prior information.

Apart from the techniques employed in the literature, large language models (LLMs) have gained popularity lately. Although mixed-language training data is used to train many LLMs, English remains the preferred language [65]. Multilingual data is used to enable LLMs to process inputs and generate responses in multiple languages. LLMs are capable of doing effectively in translation even when they are not specifically trained for such tasks. There are studies in the literature where LLMs with known success such as ChatGPT, GPT-4, etc. are used for MT tasks [66], [67]. Some studies in the literature have found that when LLMs are used for the translation of low-resource languages, they underperform the models with the best results so far [66]. In addition, studies show that LLMs achieve impressive results when translating in the XX-English direction, but relatively poor results in the English-XX direction [65], [66]. Even while LLMs work effectively on a variety of translation tasks, low-resource languages and the English-XX translation direction still need work.

C. OTHER SURVEY STUDIES

This section is a review of other survey/review studies in the literature. Some of the studies that have been carried out to date and information about the characteristics of these studies are given in Table 1. When studies are examined, it is seen that the reviews are generally studies in the field of NMT regardless of the scenario (low-high), and especially in recent years, the number of survey/review studies on low-resource scenarios has started to increase. To the best of our knowledge, there is no systematic literature review in low-resource NMT. Unlike most survey/review studies, our study covers only the low-resource NMT area. The differences between our review from other studies are as follows:

- As the study is a systematic review, how the reviewed studies were obtained is shared.
- The reviewed studies were categorized in terms of the areas they focused on.
- The most preferred methods in the studies were identified.
- The language pairs most frequently studied in the low-resource NMT literature were examined.
- Bilingual and monolingual corpora used in the studies were examined.
- The metrics used to measure the success of the studies were analyzed.
- The development tools used in low-resource NMT studies were examined.

III. METHODOLOGY

In this study, a systematic literature review was conducted for the field of low-resource NMT. While conducting the study, the process was divided into several stages. The stages of the study are given in Figure 4. In the rest of this section, these steps are explained in detail.

A. RESEARCH QUESTIONS

This SLR aims to study the deep learning techniques and applications used in low-resource NMT from 2018 to 2023 (inclusive). With this purpose, the following three research questions (RQs) are aimed to be answered:

RQ1: What is the focus of work in low-resource NMT and on which language pairs are studies conducted?

RQ2: Which deep learning methods are preferred in low-resource NMT and which evaluation criteria are used?

RQ3: What are the corpora and development tools used in the studies?

B. SEARCH METHOD

The collection of sources for this study was done through seven different databases; IEEE Xplore, ScienceDirect, Scopus, Taylor Francis, Web of Science (WOS), Wiley Online Library, and ACM Digital Library. These databases are frequently used in systematic literature searches in the field of engineering and provide a great convenience in terms of having automated search tools. Scientific study research was conducted on these databases according to the following procedure:

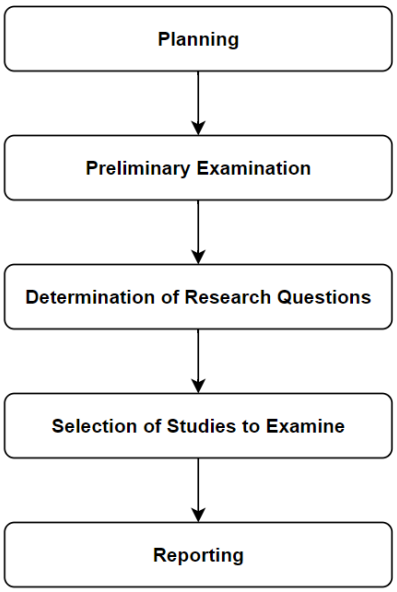


FIGURE 4. Steps followed for SLR study.

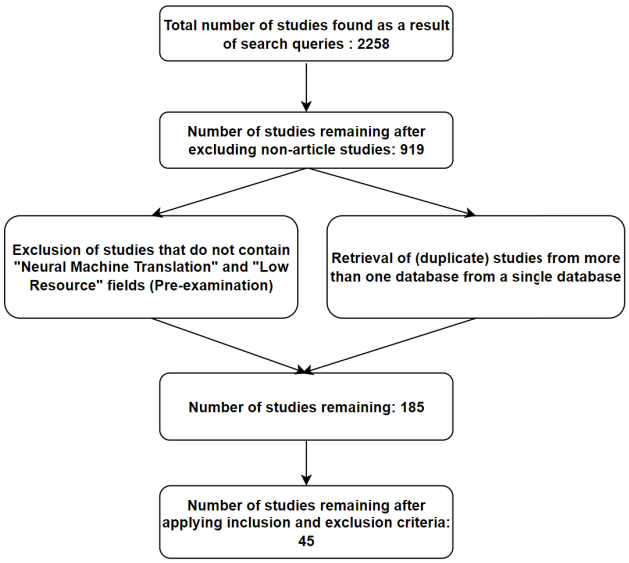


FIGURE 5. Number of studies to be reviewed for SLR study.

- With the emergence of the research questions, the searches were focused on the keywords ''neural machine translation'' and ''low resource''.
- The keywords ''low resource'' are used together with the keyword ''neural machine translation'' to focus on low-resource scenarios. Keywords such as ''transfer learning'', ''pivot translation'', ''pre-training'' and ''multilingual translation'' were created in order to reach the studies with different methods used in the field of low-resource NMT, and these words were used additionally.
- Logical operators were used to search databases. ''OR'' operators were used for synonym keywords, and ''AND'' operators were used to combine keywords.

Table 2 shows the queries used to search databases. The query for the ScienceDirect database is shorter than the others because there is a limit of eight logical operators for the queries to be used. On other databases such as Google Scholar, Springer, etc., queries, as written in Table 2, could not be written. Even if they were written, meaningful results could not be obtained (too many irrelevant results, too many studies to be analyzed). For this reason, only the seven databases from which results could be obtained were used for the study. As a result of these queries, between 2018 and 2023, 94 studies were found in IEEE Xplore, 298 in ScienceDirect, 821 in Scopus, 47 in Taylor Francis, 409 in Web Of Science, 47 in Wiley Online Library, and 542 in ACM Digital Library (as of the search date). Table 3 gives numerical information about these databases.

C. STUDY SELECTION

Although 2258 studies were found as a result of the searches, most of these studies were out of scope. In some of the databases, many studies are unrelated to the subject because the search was done for all studies. In addition, some studies may appear in more than one database in the search results. Before starting the study selection process, such duplicate studies were organized to be taken from a single database. Subsequently, some inclusion and exclusion criteria were determined in order to include studies that are appropriate for the purpose of this study. Table 4 shows these inclusion and exclusion criteria.

In the preliminary examination phase, summaries and general outlines of the studies were mainly analyzed. In this section, firstly, it was examined whether the study was in the field of NMT and on low-resource languages. Studies that did not include a low-resource setting were excluded from the review. Subsequently, 45 studies were selected to be examined according to the criteria determined from the remaining studies. Information about this selection process is given in Fig 5.

Table 5 shows the studies selected for review due to the above steps. As can be seen in Table 5, all of the selected studies were published between 2018-2023(July). From this point of view, the review we have conducted is up-to-date.

IV. OBTAINED FINDINGS AND RESULTS

In this part of the study, the studies selected for review are briefly mentioned. Subsequently, the answers to the research questions will be shared.

A. STUDY SUMMARIES

In [68], using phrase-based methods, namely phrase-based statistical MT (PBSMT) and NMT, for English to Mizo translation is investigated. The proposed model is a three-stage process of obtaining translation predictions, data preprocessing, system training and testing. The NMT system consists

TABLE 2. Queries used in databases for SLR work.

Database	Typed Query
IEEE Xplore	(("neural machine translation" OR "low resource neural machine translation" OR "NMT" OR "low resource NMT") AND ("low resource" OR "low resource languages" OR "low resource language" OR "LRL" OR "zero shot" OR "pivot translation" OR "multilingual translation" OR "pre training" OR "language model pre training" OR "transfer learning" OR "multi modal translation"))
ScienceDirect	(("neural machine translation" OR "low resource neural machine translation") AND ("low resource languages" OR "zero shot" OR "pivot translation" OR "multilingual translation" OR "pre training" OR "transfer learning" OR "multi modal translation"))
Scopus	TITLE-ABS-KEY(("neural machine translation" OR "low resource neural machine translation" OR "NMT" OR "low resource NMT") AND ("low resource" OR "low resource languages" OR "low resource language" OR "LRL" OR "zero shot" OR "pivot translation" OR "multilingual translation" OR "pre training" OR "language model pre training" OR "transfer learning" OR "multi modal translation")) AND PUBYEAR >2017
Taylor Francis	(("neural machine translation" OR "low resource neural machine translation" OR "NMT" OR "low resource NMT") AND ("low resource" OR "low resource languages" OR "low resource language" OR "LRL" OR "zero shot" OR "pivot translation" OR "multilingual translation" OR "pre training" OR "language model pre training" OR "transfer learning" OR "multi modal translation"))
Web of Science	TS = (("neural machine translation" OR "low resource neural machine translation" OR "NMT" OR "low resource NMT") AND ("low resource" OR "low resource languages" OR "low resource language" OR "LRL" OR "zero shot" OR "pivot translation" OR "multilingual translation" OR "pre training" OR "language model pre training" OR "transfer learning" OR "multi modal translation"))
Wiley Online Library	(("neural machine translation" OR "low resource neural machine translation" OR "NMT" OR "low resource NMT") AND ("low resource" OR "low resource languages" OR "low resource language" OR "LRL" OR "zero shot" OR "pivot translation" OR "multilingual translation" OR "pre training" OR "language model pre training" OR "transfer learning" OR "multi modal translation"))
ACM Digital Library	(("neural machine translation" OR "low resource neural machine translation" OR "NMT" OR "low resource NMT") AND ("low resource" OR "low resource languages" OR "low resource language" OR "LRL" OR "zero shot" OR "pivot translation" OR "multilingual translation" OR "pre training" OR "language model pre training" OR "transfer learning" OR "multi modal translation"))

TABLE 3. Number of studies obtained as a result of searches in databases.

Database	Search Result	Conference	Article	Others (review, book chapter etc.)
IEEE Xplore	94	75	19	0
ScienceDirect	298	0	235	63
Scopus	821	597	170	54
Taylor Francis	47	0	36	11
Web of Science	409	261	145	3
Wiley Online Library	47	0	44	3
ACM Digital Library	542	229	270	43
Total	**2258**	**1162**	**919**	**177**

TABLE 4. Inclusion and exclusion criteria determined for the studies to be reviewed.

Inclusion Criteria	Exclusion Criteria
1. Studies in the MT field at low-resource settings. 2. Studies used deep learning methods in the field of MT. 3. The presence of at least 1 low-resource language pair in the study or the use of another language pair at low-resource settings. 4. The work includes sentence level text - text translation. 5. Articles published in peer-reviewed journals in Q1, Q2 according to Web of Science (WOS). 6. Studies published between 2018 January - 2023 July.	1. Studies in MT that do not address the low-resource problem. 2. Studies in MT that do not use deep learning. 3. Works without text - text translation. 4. Articles written in English only.

of a one-way LSTM encoder - decoder that uses the attention mechanism for translation. The findings of the study for the NMT model can be summarized as follows; (1) the NMT system attaches importance to the accuracy of the syntactic structure of the predicted translation, as it aims to produce fluent translations; (2) the NMT system pays little attention to the precision of named entities, which usually results in partly sufficient translations; (3) translations predicted by the NMT system are shorter; (4) translations predicted by NMT are of lower quality.

In [69], inspired by humans' ability to learn languages, a new hierarchical TL architecture is proposed to take full advantage of auxiliary languages by adding a middleware for low-resource languages that only have a single parallel corpus. During the training process, the three-layer architecture transfers parameters layer by layer, and fine-tuning is done at each layer. The study was carried out between the Uyghur-Chinese languages, and the Turkish language was used as an intermediate language. The study combines the advantages of high-resource language data size, syntactic information, and linguistic similarity of the intermediate language. In terms of training time and efficiency, the model is trained several steps on a high-resource language pair (English-Chinese), and the parameters are transferred to the intermediate model in the first layer. In the second layer, the model is trained using a language pair (Turkish - English), which contains an intermediate language similar to Uyghur in terms of syntax, and the parameters are fine-tuned until they converge. Finally, to start the low-resource model, the parameters of the model trained with the intermediate language pair are transferred to the sub-model, and the model is trained on the low-resource language pair (Uyghur-Chinese) until it converges. The framework of the NMT model is not changed, but instead of randomly initializing the next model, parameters are transferred from the parent model. Transformer model was used in the study. In addition, experiments were conducted on the generalization of hierarchical TL architecture to Turkish-English. The results confirmed that the proposed method

TABLE 5. Studies selected for review after applying the inclusion and exclusion criteria.

Refs.	Paper Name	Publication Year
[68]	English-Mizo Machine Translation using neural and statistical approaches	2018
[69]	Hierarchical Transfer Learning Architecture for Low-Resource Neural Machine Translation	2019
[70]	Korean-Vietnamese Neural Machine Translation System With Korean Morphological Analysis and Word Sense Disambiguation	2019
[71]	Assembling translations from multi-engine machine translation outputs	2019
[72]	Improving Low-Resource Neural Machine Translation With Teacher-Free Knowledge Distillation	2020
[73]	Improving neural machine translation for low-resource Indian languages using rule-based feature extraction	2020
[74]	Improving neural machine translation with sentence alignment learning	2020
[75]	Multilingual Denoising Pre-training for Neural Machine Translation	2020
[76]	Neural machine translation of low-resource languages using SMT phrase pair injection	2020
[77]	Revisiting Back-Translation for Low-Resource Machine Translation Between Chinese and Vietnamese	2020
[78]	UPC: An Open Word-Sense Annotated Parallel Corpora for Machine Translation Study	2020
[79]	Pseudotext Injection and Advance Filtering of Low-Resource Corpus for Neural Machine Translation	2021
[80]	BERT-JAM: Maximizing the utilization of BERT for neural machine translation	2021
[81]	Data augmentation for low-resource languages NMT guided by constrained sampling	2021
[82]	Factors Behind the Effectiveness of an Unsupervised Neural Machine Translation System between Korean and Japanese	2021
[83]	Improving the Performance of Vietnamese-Korean Neural Machine Translation with Contextual Embedding	2021
[84]	Machine Translation in Low-Resource Languages by an Adversarial Neural Network	2021
[85]	Adaptive Adapters: An Efficient Way to Incorporate BERT Into Neural Machine Translation	2021
[86]	Enhancing low-resource neural machine translation with syntax-graph guided self-attention	2022
[87]	Addressing domain shift in neural machine translation via reinforcement learning	2022
[88]	An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation	2022
[89]	An empirical study of low resource neural machine translation of manipuri in multilingual settings	2022
[90]	Generative Adversarial Neural Machine Translation for Phonetic Languages via Reinforcement Learning	2022
[91]	GTRANS: Grouping and Fusing Transformer Layers for Neural Machine Translation	2022
[92]	Improving Neural Machine Translation for Low Resource Algerian Dialect by Transductive Transfer Learning Strategy	2022
[93]	Enriching the Transfer Learning with Pre-Trained Lexicon Embedding for Low-Resource Neural Machine Translation	2022
[94]	Framework for Handling Rare Word Problems in Neural Machine Translation System Using Multi-Word Expressions	2022
[95]	Improving neural machine translation with POS-tag features for low-resource language pairs	2022
[96]	Fully Attentional Network for Low-Resource Academic Machine Translation and Post Editing	2022
[97]	Synchronous Inference for Multilingual Neural Machine Translation	2022
[98]	Transformer fast gradient method with relative positional embedding: a mutual translation model between English and Chinese	2022
[99]	Low resource machine translation of english-manipuri: A semi-supervised approach	2022
[100]	On the Use of Morpho-Syntactic Description Tags in Neural Machine Translation with Small and Large Training Corpora	2022
[101]	Regressing Word and Sentence Embeddings for Low-Resource Neural Machine Translation	2022
[102]	Robust Data Augmenation for Neural Machine Translation through EVALNET	2022
[103]	TLSPG: Transfer learning-based semi-supervised pseudo-corpus generation approach for zero-shot translation	2022
[104]	Transfer learning based on lexical constraint mechanishm in low-resource machine translation	2022
[105]	Video-guided machine translation via dual-level back-translation	2022
[106]	A Smaller and Better Word Embedding for Neural Machine Translation	2023
[107]	Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data	2023
[108]	Neural Machine Translation of Electrical Engineering Based on Vector Fusion	2023
[109]	The neural machine translation models for the low-resource Kazakh-English language pair	2023
[110]	A Scenario-Generic Neural Machine Translation Data Augmentation Method	2023
[111]	Moprhology & word sense disambiguation embedded multimodal neural machine translation system between Sanskrit and Malayalam	2023
[112]	English-Assamese neural machine translation using prior alignment and pre-trained language model	2023

performs faster convergence and can initialize parameters for the low-resource language pair more successfully than random initialization.

In [70], the problems encountered in building a high-quality Korean-Vietnamese NMT system are identified, and solutions are proposed to address these problems. In order to create NMT system, a parallel Korean-Vietnamese corpus containing 454,751 sentence pairs was created. NMT systems are built based on attention-based seq-to-seq architecture. The experimental findings demonstrate numerous advantages over current MT systems that employ statistical and neural techniques. In addition, the Korean word sense disambiguation (WSD) method was proposed based on UWordMap, a manually constructed lexical semantic network (LSN) for particular features of Korean. WSD and morphological analysis are applied to Korean texts in the corpus. Morphological analysis segments each Korean word into morphemes, and their original form is recovered. Morpheme segmentation increases token size, and recovery of original forms reduces word size. WSD operation increased the vocabulary by labeling different meaning codes in the same word form. The Vietnamese texts in the corpus were used for word segmentation with the RDRsegmenter. RDRsegmenter reduces token size and expands vocabulary by combining tokens into a single word. In the study, the encoder-decoder structure of NMT systems is constructed utilizing deep multi-layer LSTM networks. The extracted linguistic features are used individually and in combination with the models. The best results were obtained when UTagger and RDRsegmenter were used together.

In [71], a system combination model is proposed based on the idea that an increase in the accuracy of translation

systems can be achieved by combining the outputs of SMT and NMT systems. To improve accuracy, several machine translation outputs are combined with the system combination model (SCM). Additionally, it's possible that some translated output components produced by one system will be better to their corresponding components produced by another system. A SCM can be used to obtain the benefits of both systems. SCM can be classified as either statistical-based (SBSC) or neural-network based (NBSC). However, both methods have their own advantages and disadvantages. In this study, in order to combine the outputs of different SCMs, a coupling-based hybrid architecture consisting of both statistical and neural network-based coupling techniques is proposed. The aim of the study is to gain the advantages of various MT systems and system combination models by using the translations created without knowing their detailed architectures. The proposed architecture works as three layers. First, n candidate translations are generated from N systems whose internal structure is unknown. The statistical and the neural network-based approach are then used to merge the results from the first layer in the second layer. Finally, the suggested hybrid model for the system combination chooses the best sentences produced by the different SCMs. In the study, the BiLSTM-attention model was used in the neural network-based approach. The outputs of four systems (phrase-based, Hiero, NMT, Google) from existing studies in the literature were combined. These models are trained with different corpora within themselves. The sum of the separate elements complexity makes up the complexity of the hybrid SCM. It has been observed that the method used affects the overall working speed but shows significant improvements in translation success. Phrase-based, Hiero, and Google were used in combination to get the best outcomes.

In [72], a new Teacher-Free Knowledge Distillation framework is proposed. Transferring knowledge from one neural network (teacher) to a different one (student) is the goal of knowledge distillation (KD). A target distribution that looks like a virtual teacher model is manually created in the study. The target distribution depends on how many terms in the target vocabulary are similar to each other. The loss function in the MT model training is increased, and the diversity in the vocabulary is modeled more accurately. To enhance model training, a further Kullback-Leibler divergence loss is applied based on the maximum likelihood estimation. Two probability distributions are compared to determine the additional loss. The model's training prediction provides the first probability distribution, while the distribution obtained through word similarity provides the second distribution. The vector representation of tokens (words/subwords) in the vocabulary was obtained from large monolingual data. Cosine similarity is used on pre-trained embeddings to rank the token order. FastText and CCMT2020 corpus are used for the pre-trained embeddings. The proposed method is compared with sequence-level knowledge distillation and Transformer model and achieves better results. In addition, the proposed system is tested together with the back-translation method. Although back-translation has additional training cost, it is found that it can further improve the effectiveness of the NMT model.

In [73], a NMT model between Sanskrit-Hindi languages is proposed by combining RNNs with a rule-based linguistic approach. In order to train and test the proposed NMT system, several models, activation functions, training data, and lengths of sentences were used. The suggested technique uses a pipeline design that accepts input from its earlier stages, performs calculations, and forwards the result to the following action. The rule-based pipeline design for the NMT system has 10 modules. To more effectively train the system, each module provides a distinct output as linguistic attributes to the encoder-decoder system. The encoder-decoder with attention is implemented as a stack of Bi-GRU layers. The proposed framework integrates features from the rule-based pipeline architecture to train the RNN. Each feature has a separate word embedding. To combine all these word embeddings, it creates a feature embedding matrix as a sum of all features embedding sizes. As the lengths match, these embeddings are subsequently added to the overall embedding size. These retrieved linguistic features are multiplied by the input vectors. Only this update to the encoder is made; all other functions and parameters remain unchanged. Initially, a small parallel corpus was used to train the NMT system. In this way, the system achieved low accuracy, and the output was not intelligible. Therefore, data augmentation techniques are included in the system to overcome this problem.

In [74], problems of inadequate translation were addressed by imposing sentence alignment constraints on NMT. The alignment score between the source and target sentences is predicted using a discriminator (D) based on sentence alignment. A gated self-attention based encoder is used in D to capture evidence of semantic alignment of input data. In order to avoid over-penalizing for translations that are correct but not human-generated, the N-pair loss is defined in the training process of D. Then, an adversarial training and alignment-based decoding strategy was applied to integrate the sentence alignment constraint into the NMT. A basic NMT model is trained using adversarial training to create accurate translations that outperform those produced by the generator (G) and discriminator (D). D guides the NMT model for alignment-sensitive decoding by integrating the alignment score and decoding probabilities when generating a translation. The proposed encoder is a structure that learns to focus on lexical information important for sentence alignment and to improve the contribution of keywords. This semantic and lexical information is transferred to the NMT with the suggested training and decoding processes. The alignment-sensitive decoding structure allows the decoder to consider adequacy and fluency of translations. These features incentivize the NMT model to generate translations that match the semantic information a discriminator learns for sentence alignment. In the study, the LSTM-attention and

Transformer models are implemented.Uyghur-Chinese were the low-resource language pair employed in the study.

In [75], it is shown that substantial performance improvements can be achieved by pre-training an auto-regressive model with a target that extracts and reconstructs noise from texts in several languages. In this study, a multi-lingual seq-to-seq mBART model is presented, which de-noises the autoencoder. BART is used to train mBART on sizable monolingual corpora across many languages. Noise is created in the texts by masking the entered texts and replacing the words. A single Transformer model is trained to recover these texts. Unlike other NMT pre-training methods, mBART pre-trains a full auto-regressive seq-to-seq model. Without any task- or language-specific modifications or initialization procedures, mBART is trained once across all languages, producing a set of parameters that may be fine-tuned for each of the language pairs, in both supervised and unsupervised circumstances. Although BART only received pre-training for English, pre-training influence on several language pairs have been systematically studied. To more accurately assess the effects of various levels of multi-lingualism throughout pre-training, models using all languages and pre-training with fewer languages were created. The training data was divided into high, medium, and low-resource scenarios, and various experiments were conducted. mBART pre-training has been shown to provide constant improvements in performance at low/medium-resource settings and outperform other existing pre-training schemes against bilingual models and BT. It has been found that mBART can boost performance even for languages that are not included in the pre-training corpora.

In [76], a method is proposed to enhance the NMT system in languages or domains with limited resources. In the study, phrases taken from an SMT system are used as training data for NMT. The basic idea of this method is to supply more details about the compatibility between source and target expressions. A sentence pair does not contain any information regarding the mapping between the source and target expressions when it passes through an encoder-decoder. The model learns translation maps implicitly by predicting and correcting the error over a across a vast parallel data. However, the model cannot comprehend the relationship between expressions when the amount of data is small. Therefore, in addition to feeding sentence pairs to the network, sentence pairs were also fed as training examples. To implement this feedback mechanism, sentence pairs were extracted from the original training data using the Moses SMT system, and they were added to the original training data as parallel sentence pairs. Experiments were conducted with two methods; attention-based GRU and Transformer models. The results of the proposed method were better to those of basic models, and Transformer model performing the best. In addition, the proposed approach was compared with some techniques discovered in NMT, such as sub-word level NMT and back-translation, and achieved better results.

In [77], the effects of BT on NMT were investigated using language pairs that not only utilize distinct writing systems but also belong to different language families, leading to more challenges for MT with limited resources. With models trained on extremely low-resource corpora, SMT and NMT experiments were carried out with character and word based settings, offering comparisons for Chinese-Vietnamese and Vietnamese-Chinese directions. Additional analyses, including N-gram F1 score, error rate, and linguistic analysis, were also performed to obtain new results. The study also examined impact of synthetic data size on model performance. Although different results were obtained, NMT models generally achieved better results when a large amount of synthetic data was used. When word-based SMT and word-based NMT outputs were examined, it was discovered that NMT outputs are better in two ways; a) the number of untranslated Vietnamese words is much less than SMT, including named entities; b) in NMT outputs, the word order and general syntactic structures are more precise and comprehensible. The study concludes that in the two translation directions of Chinese-Vietnamese, the addition of artificial data positively affects the performance of character and word based models. For bidirectional Chinese-Vietnamese translation, the performance of SMT outperforms NMT in most cases.

In [78], firstly, a parallel corpus called UPC, consisting of two large parallel corpora, was created to train Korean-English and Korean-Vietnamese MT models. Data was gathered on subjects that were pertinent to everyday life, such as economy, education, religion, etc., for a variety of audiences. Word ambiguities (or homographs) that have the same spelling but different meanings harms both SMT and NMT performance. This model forces NMT systems to choose from several candidate translations representing different meanings of a word. To solve this problem, a hybrid approach is proposed combining knowledge-based methods with a sub-word conditional probability to determine the suitable purposes of homographs and explain the codes corresponding to these homographs. Using this approach, a fast and accurate WSD system called UTagger has been developed. WSD was then applied to the original Korean sentences in the UPC. The SMT and NMT systems were both trained using this corpora. Experiments were carried out with the normal version of the corpus and using UTagger. Better results were obtained when using UTagger. In addition, rare words produce a large number of out-of-vocabulary (OOV) words, which is a problem for MT. TClear word boundaries were formed and OOV words were decreased by the morphological analysis used in UTagger. As a result, WSD usage has enhanced MT system performance. Additionally, the NMT model achieved better results than the SMT model for these corpora.

In [79], by repeatedly using the Transformer model for BT, it is proposed to add a pseudo-parallel corpus to the training data. A successful round-trip approach is analyzed with sentence alignment metrics for pre- and post-translation filtering. If the target sentence and the round-trip translation are parallel, the synthetic source sentence is considered a

possible match to the monolingual target sentence. Therefore, this sentence can be added to the synthetic parallel sentence corpus to increase the training efficiency. The proposed framework is composed of two modules. The first module, back-translation, consists of 4 steps. First, an iterative Transformer is trained on the source-target pseudo-parallel corpus with different parameter settings to acquire the synthetic data. Then, at each epoch during training, the translation is analyzed using the source synthetic data, and the model with the best BLEU score is selected. The Transformer model is tuned with different iteration patterns and layer sizes to reduce training variance. Then, sentence parallelism between monolingual target and source synthetic sentences is estimated. Finally, Cohen's Kappa measures the agreement between the mono-lingual target data and the synthetic source data to avoid duplicate sentences in the corpus. Sentences with low Cohen's Kappa scores are removed from the corpus as they are considered detrimental to model performance. The second module in the proposed framework is the round-trip approach used to acquire the target data. First, this module uses a round-trip translation of synthetic source sentences in the source-target direction to obtain a synthetic target sentence. Then, the similarity between monolingual and synthetic target sentences is computed, and low-scoring sentences are filtered out. Finally, the filtered synthetic and monolingual target sentences are combined to extend the training data. The proposed method is compared with some works in the literature and achieves better results in both high- and low-resource scenarios in different language pairs.

In [80], a NMT model called BERT-JAM, which stands for BERT-fused Joint-Attention, was proposed. Three ways have been tried to maximize the use of BERT for NMT. First, a fusion module is included in each encoder/decoder layer to be able to use the representations of BERT in a combined representation. Weights are shared between different samples to combine multi-layer representations. Second, it is proposed to integrate the fused BERT representation with the encoder and decoder layers by combining the self-attention and the cross-attention modules using one joint attention module. The joint attention module carries out multiple attention module tasks at once, dynamically allocating attention between BERT representation and encoder - decoder representation. A joint attention module termed BERT-encoder joint attention is employed at each encoder layer to take part in both BERT and encoder representation at the same time. Each decoder layer contains two joint attention modules. In the first, called BERT-decoder joint attention, deals with combining the BERT representation with the decoder representation. Secondly, called encoder-decoder joint attention, focuses on both encoder and decoder representation. Third, to address the issue of catastrophic forgetting, the proposed model is trained using a three-step optimization technique that gradually solves various model components. By employing this method, the model is able to benefit from the improved performance that BERT fine-tuning provides.

In [81], an easy-to-use but powerful constrained sampling technique is proposed for data augmentation (DA) in NMT. Constrained sampling method that makes use of edit distance calculation are considered to be more effective than other methods that choose words at random from the original text. The proposed way is basically similar to GAN networks and can be expressed in 3 steps. First, both positive and negative samples were used to train the discriminator sub-model. Therefore, some negative samples were created from original data using the negative sampling technique. Second, the evaluation sub-model is trained using original and generated negative data. The evaluation sub-model is designed to select high quality data after generation and is intended to ignore, to some extent, sequences containing semantic or syntactic errors. Third, some samples are augmented using the edit distance sampling technique on the original data distribution, and low-quality augmented datas are ignored by the discriminant sub-model. The proposed method stands out because it is language-independent. Such a sampling method can be incorporated into NMT systems in different languages. The proposed strategy performs noticeably better than the approaches in the literature, according to experimental findings.

In [82], investigates which knowledge a model gains from pre-training and which information from the pre-trained model enables a highly accurate unsupervised NMT. For this reason, which layers of the unsupervised NMT system store what kind of information and whether features such as word order of cross-attentions differ in languages have been analyzed. The cross-attentions of an encoder-decoder architecture are being analyzed using a novel technique that takes into account the different features of the source and target sequences. A language generation model is pre-trained using the Masked Sequence-to-Sequence (MASS) method with two monolingual corpora. The pre-trained model is then fine-tuned for the same corpus and unsupervised NMT task with a back-translation loss. The Transformer method is used for the architecture consisting of an encoder - decoder. An input sentence including a random masked fragment is provided by the encoder, and the decoder attempts to predict this random masked fragment. In this work, a BT approach is adopted to build an unsupervised NMT system because the BT approach can be easily implemented by using a typical encoder-decoder for both languages with the MASS method. Both strategies rely on the creation of cross-lingual word embeddings across the two languages before an unsupervised NMT system is trained. The results show that pre-trained models are helpful in improving the performance of a unsupervised NMT system.

In [83], the benefits of adding linguistic annotation to sentences used as input for MT are investigated. A model for Korean-Vietnamese NMT is proposed that combines a Transformer model with a pre-trained Viet-BERT model. The Korean-Vietnamese bilingual corpus undergoes a number of pre-processing processes before being incorporated into NMT systems in order to enhance the standard of NMT.

Especially, POS-tags have been added to Vietnamese sentences, and morphological analysis (MA) and WSD have been applied to Korean sentences. A BERT-based model for Vietnamese sentences was applied to the encoder layer to create an embedding of each token of the given input. To compare the effectiveness of proposed NMT technique, different MT systems for Vietnamese to Korean have been created with varying formats of input. The most important is the BERT fused Transformer model, in which the BERT-based VietBERT model is used. BERT can identify what a word means based on context and produce relevant embeddings for different contexts, in contrast to context-free methods like word2vec. Additionally, the outputs of the BERT model are maximally utilized. This improved representation is then connected to each layer of the NMT model via the attention mechanishm. Consequently, due to this input, the decoder of the NMT model produces more proper target sentences. Using BERT also improved POS tagging results, an annotation of Vietnamese data. As a result, combining the VietBERT and NMT increases the success of Vietnamese-Korean MT. Other models created are Bi-RNN encoder-decoder-based NMT models.

In [84], a study on adversarial learning is presented. To get correct translations in complex systems, an enhanced feature extraction method is examined in small-sized training of sentence pairs. The suggested model additionally makes use of TL to further improve NMT performance. The whole GAN system consists of a discriminator D and a generator G. The parameters of G and D are optimized using two adversarial losses. G uses fake examples to perplex D, the discriminator. Conversely, discriminator D seeks to identify the fake examples produced by G and adjusts its parameters as necessary for this. The adversarial losses of a GAN model are included in the NMT as they help the LRL translation. In the proposed model, RNNSearch is designed in the generator part, and a residual connected convolutional neural network (CNN) is designed in the discriminator part to classify the input pairs according to their hierarchical features. Mixture, Res, and Feature are the three basic components that make up the discriminator.The two different embeddings in the input pair are independently sent through an exclusive convolutional layer and merged in a mixture block. This block contains an ordered convolutional layer to thoroughly fuse dense exponential linear units (ELU) and their embeddings. Res block combines the same number of layers under its predecessor faster. By contrasting the suggested model with other models, its efficacy has been confirmed. These are models such as RNNSearch, BERT, and ALBERT. Next, the pre-transfer trained generator, discriminator, generator-discriminator, and non-transfer training models were compared. The analysis of the proposed model in terms of TL is tested with a separate generator, a discriminator, and both a generator and a discriminator. The best results are obtained when only the discriminator is transferred.

In [85], it is recommended to use pre-trained BERT model on both the encoder and decoder sides to improve the use of information obtained through pre-training and to support performance. The study used lightweight neural network components called adapters to incorporate the BERT model into the seq-to-seq framework. Two pre-trained BERT models are added on the source and target side, and these are considered encoder/decoder. The advantage of adapter modules in the model is that they are parameter efficient and robust. The suggested system also doubles the decoding speed through parallel decoding. Each element in the proposed structure can be thought of as a plug-in unit, which makes the model very flexible and task-independent. Since pre-trained BERT models are deep models, this study investigated whether it is necessary to add adapters to each BERT layer. A probabilistic learning process is utilized to determine whether to use an adaptor in each layer using hidden variables. Variational inference optimizes the latent variables, and an additional loss function regulates the number of adapter layers. In this way, the parameter scale of the adapters is automatically pruned, and the adapter layers are directly fine-tuned to a pre-trained model, which significantly lowers the model decoding delay during inference. Based on the fact that some layers in the Transformer model can be pruned without seriously harming the model effectiveness, it is also assumed that some adapter layers can be pruned as well, as not all adapters play essential roles when fine-tuning. In order to enable the model to choose and employ adapter layers automatically, a probabilistic approach is applied. The proposed model performed better when compared to some research in the literature.

In [86], a syntax-graph guided self-attention (SGSA) tecchnique is proposed, a model that combines the syntax of source sentences with stacked multi-head self-attention layers, aiming to improve Transformer by using syntax more instantly. The syntax is converted to a graph in order to create an effective combination with the NMT model. The syntax-sensitive approach is a structure suitable for sub-word units, and it resolves the issues caused by extensive vocabularies and sparse words. The source-side syntactic dependency is used as a guide, and a syntactically directed self-attention mechanism is used without additional parameters. To perform this process, which the authors call dynamic multiple syntax-aware self-attention representations (DMSR), the syntactic graph is adaptively tuned, and the effect of different fusion methods on the performance of the model is investigated. To solve the absence of syntactic information and maintain the parallel computational ability of self-attention networks, the syntactic relationships of each source token are represented as vectors and applied to the self-attention components Query-Q and Key-K. Different strategies have been attempted to integrate DMSRs and attain the ultimate representation. These strategies encompass methods such as average-pooling, highway and linear networks, all used as fusion methods. The analysis showed that adding syntax information in the first three layers of the Transformer decoder yielded better results, while adding it in the deeper layers did not result in significant improvement.

When syntactic information is integrated into a single layer, performance often degrades with increasing layer depth. In addition, the proposed model was compared with other models in terms of inference time and parameter count. The proposed method has fewer parameters in all translation directions.

In [87], an approach is proposed that strengthens the relationship between languages to improve translation quality and addresses domain adoption issues through reward-based learning. To address the domain adoption problem, a Reinforce-based Sentence Selection and Weighting (RSSW) method is proposed that chooses data based on the rewards received. After training the NMT model on out-of-domain data using RSSW, the NMT model is fine-tuned on the in-domain corpus using maximum likelihood estimation and minimum risk training. The three modules that make up RSSW are translation model training, policy network, and language model. The first sentence weight for each training sentence is determined using the LM. Then, the sentence weight modification module adjusts the sentence weights in accordance with the action values, meanwhile the reinforcement learning (RL) agent assists in generating change actions according to the environment states through the policy network. In the final, NMT training is carried out using weighted out-of-domain training sentence pairs and fine-tuned on the original in-domain data. Both the LM and NMT model are trained using a Transformer architecture. The proposed model has been compared with studies in the literature by only in-domain, in- and out-domain together, in-domain training - out-domain fine-tuning, and got better outcomes. In addition to these, LSTM architecture was also tried instead of Transformer architecture. RSSW showed $\sim$1 BLEU score improvement in the LSTM model. In addition to the original target languages, German, English, and French were used in the experiments in low-resource scenarios, and the model performance improved by $\sim$1 BLEU point.

In [88], a novel technique for producing synthetic data in low-resource scenarios is proposed, and compared to BT in experiments. The suggested approach is quick, reliable, and does not need any additional outside resources, such as dictionaries, pre-trained models, rules, or language models. No additional resources are used for data augmentation except for a small amount of available bilingual corpora. It was suggested that artificial translation units (ATUs) be used for data augmentation while maintaining the sentence word order and context. ATUs refer to tags produced by standard translation modules using solely monolingual vocabulary.Based on ATUs, equivalent artificial sentences from the target are created for an original sentence pair. The original source sentence is then compared to these pseudo-target sentences to produce artificial data. The artificial corpus is mixed with the original corpus to train the NMT systems. To enhance the effectiveness of the MT, the method can be utilized in conjunction with the BT. In addition to data augmentation, the effectiveness of low-resource translation is investigated in terms of combining Chinese and Japanese texts on the source side (combined training) in a translation task to Vietnamese. The translation tasks capitalize on the utilization of shared translation units between the two languages. Furthermore, the BERT model is incorporated into NMT systems. The BERT model is trained using a mixture of Chinese and Japanese texts (with grammatically inverted patterns), in contrast to previous studies. The goal is to look at how effective combined training systems are in high-resource scenarios.

In [89], highlights the low-resource MT, specifically for Manipuri as well as other Indian languages. This is achieved through a multi-lingual technique and involves direct translation scenarios between Indian languages and Manipuri using zero-shot manner. It has been mentioned that there is a capacity bottleneck issue in a single shared MNMT model. In order to address this, a comprehensive analysis is performed on the multi-lingual cross-lingual word embedding (MCLWE), which precedes the MNMT model. It has been shown that this increases the generalizability of the model. In addition, the effects of using such an embedding on zero-shot translation are also examined. In the proposed method, firstly, embedding training is performed for all languages separately. The process of multi-lingual alignment is then carried out by mapping each language embedding to the shared language embedding area. In the study, a single common MNMT model is used in a many-to-many manner with shared encoders and decoders. The encoder-decoder are initialized, and the multi-lingual model trains together over N language pairs. The proposed method is compared with bilingual, multi-lingual, and pivot translation models. In the test processes, it was observed that more improvement was achieved in the English-XX direction than in the XX-English direction, with the inclusion of MCLWE in the system for all language pairs. Finally, the model was compared with the pivot-based methods using zero-shot translation, and competitive results were obtained, which were not better than the pivot-based cases. Overall, the proposed method can handle repeating words better than the bilingual and multi-lingual base models and enhances the quality of NMT for the low-resource Manipuri language.

In [90], a study was conducted to reduce corpus requirements and improve context learning in extremely LRLs. A new method is proposed that jointly embeds textual and phonetic information of languages into GAN-NMT by leveraging an optimized attention network based on deep RL. In the proposed architecture, a pre-trained NMT model is utilized as the generator, creating translations from the source sentences, while a different network is employed as the discriminator, determining if the translations are genuine or not. A new GAN model consisting of Deep-RL-Guided-Attention as the generator and a Convolution Neural Network (CNN) as the discriminator is used to obtain better attention weights. The Transformer model has been modified to create the generator model. The discriminator is a classification model developed using CNN to distinguish between real sentences and sentences generated by the generator. Instead

of using the conventional word embedding, the GAN model is trained using a new joint embedding. By substituting deep RL-guided attention for the initially used attention in the generator, the suggested design enhances the GAN model and raises the probability of deceiving the discriminator model. The enhanced GAN model can learn additional phonetic context that is missing from other approaches by combining information from textual and phonological representations. The proposed method is compared with different approaches in the literature and achieves better results. In addition, the proposed framework has been tested on high-resource German-English translation and outperforms some of the compared models.

In [91], a method called Group-Transformer (GTRANS) is proposed, which strengthens the model and separates multiple layers into different groups to take full advantage of the low-level and high-level attributes in both the decoder and the encoder. Only the last latent states from each encoder group, which is composed of a certain number of contiguous layers, are included in the combined representation. Similar to this, each decoder layer is broken up into distinct decoder groups before being integrated as a whole. The target words are predicted based on the word probabilities generated by combining all of the representations of each decoder group, allowing low-level information to also directly influence the predictions. Experiments were done with 5 different corpora in the study, but the corpus that can be called low-resource is only the IWSLT-14 English-German. The proposed approach was contrasted with several approaches suggested in the literature, and the findings were presented. The technique provides $+0.78$ and $+1.73$ BLEU score enhancements in De-En and En-De directions, accordingly, indicating that the proposed method can take advantage of multi-layered features to improve translation quality significantly. In addition, the IWSLT-17 corpora was used for multi-lingual experiments in the study. The authors said that the different layer representations provided by the proposed method are suitable for the multi-lingual translation task and provide consistent improvements in all translation directions of the model. In comparison with the Transformer model, the proposed method has no additional parameters and has a close inference speed.

In [92], a NMT model trained on a sizable corpus including every Arabic dialects was created. The goal is for this NMT model to be able to translate a particular dialect using a low size corpus. A transductive TL strategy is proposed to address the issue of data scarcity in Arabic dialect translation. The transductive TL strategy was used with two NMT models: LSTM seq-to-seq and attention seq-to-seq (Luong attention). The corpora used in the proposed framework are MADAR (25 Arabic dialects-Modern Standard Arabic (MSA)) and the target PADIC (Algerian dialect-MSA). The suggested strategy consists of two key steps: 1. learning step: the LSTM seq-to-seq and attention seq-to-seq models are trained using the 25 Arabic dialects of the MADAR corpora; TL step: using TL, the LSTM seq-to-seq and attention seq-to-seq models are trained again using the Algerian Arabic dialects in the PADIC corpora. The parent model is built using the massively parallel corpus MADAR. The parent model is then retrained using the PADIC corpus, which results in the creation of a child model. The information from the parent model is passed down to the child model through the reuse of its parameters. The model performances started dropping for sentences with more than 20 words for the LSTM model and more than 25 words for the attention model. In addition, the study compared the results with studies using different TL methods, but this comparison is improper since the corpora in the studies are different.

In [93], proposed model overcomes the drawback that TL does not take into account the vocabulary properties shared by the parent and child models when fine-tuning the sub-model. Based on this situation, this study proposes a method to use vocabulary embedding and vocabulary information in the child model. The integrated corpus used with the shared vocabulary while training the parent model in the proposed structure shows a much better translation performance by using the parent and child vocabulary. To make the model strong, the data used in the child model has been added to the parent model. In this way, the parent model has prior knowledge of the child model. Therefore, It is thought that a more robust TL approach will emerge. The basic idea is to share vocabulary properties between the parent and child models before fine-tuning. In the suggested structure, the parent model is first trained using a language pair with a lot of resources. Afterward, preparation is made for the hybrid model in order to create an integrated corpus. The oversampling method is used to create the integrated corpus. A larger mixed corpus is produced by bringing the size of the child corpus to parity with that of the parent corpus. A joint vocabulary is created over this corpus, and the hybrid model is trained. The child model is fine-tuned over this hybrid model using a low-resource language pair. Languages used for parent models are Arabic, Persian, and Turkish. The suggested model is built upon the Transformer method. The proposed strategy performed better when compared to several TL techniques in the literature. Additionally, an experiment was conducted on the model using a low-resource (Uyghur) language in the parent model, and it was seen that the results improved.

In [94], a study was conducted on how to handle with out-of-vocabulary (OOV) words and multi-word expressions (MWE) in a NMT system. NMT systems use the softmax function in the output layers. Softmax function has a high computational complexity, and therefore NMT systems are used with limited vocabulary sizes. This situation triggers the OOV problem. MWEs are constructions that contain multiple words but behave as a single word. NMT systems may fail to learn, remember and reproduce MWEs as they represent the entire sentence in a high-dimensional vector. The Punjabi-English language pair is analyzed in this study, and existing systems in the literature are studied. In addition, a corpus of MWEs and named entities were created. In the study, the encoder-attention-decoder structure was used with

LSTM, and a total of 4 different models were analyzed within this structure using word embedding and different corpora.Word vectors obtained from FastText were used in the study. In addition, a pre-processing module was created for sentences. It has been observed that the models are generally more successful in short sentences, and there is a decrease in success after 15 words.

In [95], Transformer, multi-source Transformer and shared multi-source Transformer models with additional grammatical features are used for low-resource NMT. The major goal of this study is to enhance the translation efficiency of low-resource languages by including extra linguistic variables into NMT models. In the experiments, POS taggers were employed to assign the accurate POS tag to every word within the corpus. A POS labeling format, Word|POS, was used in the experiments. To implement translation models, on the source side, POS-tags are first used. In order to initiate translation models, POS-tags are then added to the target side. Then, for each translation model, POS-tags are included in every word on both the source and target side. Multi-source Transformer and shared multi-source Transformer models use two inputs (i.e., sequence data and POS-tagged sequence data), and the models outputs are either sequence data or POS-tagged sequence data. The basic model is the Transformer model, which uses only word vectors. The multi-source Transformer model is an enhanced version of the Transformer. This enhancement involves incorporating an extra encoder and adding an additional target-source multi-head attention component on top of the existing one. This modification allows for the utilization of double inputs same time. The architecture includes two encoders, one for the words and the other for the linguistic features. Despite the fact that two independent encoders use the same parameters, their outputs differ, and they merge in different spots throughout the decoder. Shared multi-source Transformer and multi-source Transformer models have many similarities. In contrast, when training, the parameters of the multi-source Transformer model are shared. In addition, the authors proposed a POS Tagging method and included it in the experiments. With this method, competitive results were obtained with fewer labels. Generally, the best results were obtained with the shared multi-source Transformer method.

In [96], the objective is to enhance the translation success by combining two Transformer-based structures on the Turkish-English language pair with shallow fusion method. Firstly, a Turkish-English corpora was created for study. Transformer and SciBERT models are used together in the proposed structure. To maximize the effectiveness of these two architectures, the shallow fusion technique is used. The outputs from the decoder and LM are combined in another neural network (NN). The additional NN structure used in the study is the fully attentional network. On the fusion process, the combined output goes through token, positional, and segment embedding. After these processes, the output is given as input to the Transformer encoder. Meanwhile, the weights of the LM are frozen. The proposed model has been compared with Google Translate, LSTM, and Convolutional Based Transformer and received better results. Additionally, the proposed model is tested on the WMT'17 and WMT'18 corpora in a zero-shot manner. 20.12 and 20.56 BLEU scores were achieved in the Turkish-English direction, respectively, and although the results are not very good, they are considered to be competitive.

In [97], a fully synchronized inference technique is proposed for multi-lingual NMT, which can simultaneously and interactively produce multiple target sentences in several languages.In the inference phase, the model uses predicted words in other languages additionally the source and the prior predicted words while predicting the next word. A module called cross-lingual attention has been developed that can dynamically choose the most pertinent part from the target sentences of more than one language in order to utilize the supplementary information of different target languages during generation and to direct the generation of the language of interest. This allows the approach to generate translations in multiple languages at the same time, and this allows for mutual enhancement between target languages. The study is built upon the Transformer model. In the proposed model, the encoder component is identical to the original Transformer model. In the decoder part, the recommended cross-lingual attention module is used. This method generates a simultaneous representation for each language. In this way, the attention calculation for a language pair establishes a relationship not only within itself but also with other languages. For this, the attention between the two languages is calculated first. Afterward, these binary attentions are merged using a fusion function to form the ultimate representation. Three different fusion methods; linear, non-linear, and attention-based, were used in the study. While linear and non-linear methods behave equally for target languages, the attention-based method is designed as a structure that allows dynamically selecting relevant information from all languages. The beam-search algorithm has been modified to make inferences in more than one language in order to be suitable for synchronous inference. This enables interaction between different languages throughout the decoding process. Model training was conducted in the form of multi-task learning to take advantage of existing large bilingual parallel data. The corpora considered as low-resource in the study is IWSLT'14. In total, 5 different situations were tested. These; standard bilingual Transformer, fully parameter sharing multi-lingual Transformer and fusion methods are the case. In the study, better results were obtained by adding a small amount of parameter compared to the multi-lingual shared model. Looking at the results, it was seen that Chinese is the language that contributes the most to translation among languages in the cross-lingual attention. In addition, the proposed method has been tried by using different languages as a structure, and better results have been obtained compared to the multi-lingual Transformer model.

In [98], the authors mention two problems with the Transformer model. The first one is that when positional encoding

is done on the corpus, the location information is lost, and the model entirely disregards the sequence order. Second, there is often an over/under-translation problem, and the model does not capture the correlation between words well. To overcome these problems, the Transformer fast gradient with relative positional embedding (TF-RPE) method is proposed together with the adversarial training method. The proposed method can get local and global interdependencies among texts by replacing absolute positional encoding with relative one. To improve the training of word vectors in the multi-head attention part of the Transformer, the fast gradient method (FGM) adversarial training algorithm is added. In the proposed structure, words are first converted into word vectors in the embedding layer. Then, to obtain the desired positional embedding information, the location information formula is used to add the location code to the word vectors at each position using RPE. Obtained results are transmitted to the encoder and decoder sections of the Transformer model for training, and the FGM adjust the training data for the encoder layer. The FGM adversarial training algorithm is added to the attention module to enable the model to identify more adversarial examples and reduce over/under-translation problems. By adjusting parameters, adversarial training generates noise to enhance robustness and generalization. For a better adversarial sample, FGM typically utilizes a perturbation value that scales with the gradient. The loss in the multi-headed attention of the encoder layer, along with the gradient value, the embedding layer gradient, and the norm value, are employed to obtain a new loss and its gradient. The parameters are updated for improved model convergence using a combination of the initial and adversarial gradients. Using relative position embedding and adversarial training ensures the correct positioning of words during translation and using semantic information by the Transformer. Chinese-English are employed as a low-resource language pair in this study. The proposed model has been compared with CNN-based Transformer, BERT-Fused and other approaches in the literature and has achieved better results.

In [99], concentrated on examining the efficacy of unsupervised and semi-supervised methods for English-Manipuri MT using the monolingual data that was available. This study uses self-training (ST) and BT to increase little parallel data with monolingual data on the source and target sides in order to overcome the low-resource problem with a semi-supervised system. In order to increase the amount of original parallel data, ST uses a source-target MT model to translate monolingual data on the source language. Akin to this, BT creates synthetic data from target monolingual data using a trained target-source MT model. From three supervised candidate structures (SMT, LSTM, and Transformer), a thorough analysis was done to choose the basic architecture for the suggested semi-supervised MT model. The trained models were examined, and since the Transformer gave the best results, the Transformer structure was used in the proposed model. To deal with lack of the parallel data, a semi-supervised MT system that inckudes ST, forward-translation (FT), and BT is proposed. The artificial data produced during BT and FT is noisy and there are distributions of this noise. Therefore, to randomize this built-in noise, some perturbation in the manner of word shuffling, word dropout, and word spacing has been added to induce some degree of randomness in order to alter the initial noisy distribution of artificial data. In addition, cases where only BT, only ST and both were used were compared. Using the two methods together gave much better results. Increasing artificial data is only advantageous to some extent, and performance suffers when more synthetic data is added because of more noise. To examine the impact of the BLEU score for all models with varied sentence length, test sentences were grouped according to the length of the reference sentences. However, it was observed that the success decreased as the reference sentence length increased. The performance of the suggested semi-supervised approach was assessed over alternative supervised, unsupervised, and pre-trained mBART techniques, and better results were obtained with the proposed method.

In [100], it was shown that NMT systems are able to benefit from additional morphological information for translating English-Slovene. To provide a more thorough understanding of the practiablity of morpho syntactic description (MSD) tags and to integrate MSD tags, experiments were performed utilizing various training corpus sizes and methodologies. The concept behind the proposed technique is to emphasize on preparing data rather than the design of the NMT system. Labeled and lemmatized corpora were used to create five different formats from each corpus using different methods.The best results in the English - Slovene direction were obtained when words and MSD tags were used as distinct tokens in languages. Best results in the Slovene - English direction were obtained when lemmas and MSD tags used as independent tokens on the source side, and only superfical words were used on the target side. The NMT models used in the study are Transformer and LSTM. However, it is not clearly stated in the study which results were obtained with which model.

In [101], the authors proposed a new method by extending their previous work called ‘‘regressing word embeddings (ReWe)’’. During training, ReWE is incorporated as a module into the decoder of the seq-to-seq model. As a result, the model is trained to predict the following word in the translation as well as the pre-trained word embeddings. This approach has proven that pre-trained word embeddings can take advantage of embedded contextual information, especially with low/medium size corpus. The idea previously used in this study is extended to sentence embedding regression (ReSE). ReSE employs a self-attention method for every input sentence in order to understand a single, fixed-size vector at the output. Throughout the training phase, the model is trained to regress this vector towards the pre-trained word embedding of the reference sentence. Specifically, it has been proposed to jointly regress word

and sentence embeddings as a unified training modifier, and the suggested method is named ReWE+ReSE. In order to promote model regularization, the proposed ReWe model combines information from the word vector into the loss function. A ReWE block has been added to the NMT model to produce continuous vector representations as output. The ReWE block receives the hidden vector from the decoder at each decoding step and outputs a second vector of the same size with pre-trained word embeddings. In order to achieve accurate word embedding, the model is trained to regress the predicted vector. This is accomplished by employing a loss function that computes the distance between two vectors. The authors said they used cosine similarity for this loss in the previous study. ReSE and ReWE differ primarily in that ReSE predicts one regressed vector per sentence as opposed to an one regressed vector per word. The proposed approach makes use of the LSTM and Transformer models. Additionally, pre-trained FastText embeddings are used to initialize word embeddings in both models. Pre-trained USE and SBERT sentence embeddings were used in corpora where English is the target language, as they can be used as monolingual encoders. Among the models used, the LSTM model achieved better results than the Transformer model.

In [102], a data augmentation method using a BT technique for NMT, and a neural network-based data evaluator called EvalNet, is proposed. EvalNet is determine weights for training data and augmented data. In a gradient descent step, the loss values can be modified using these weights. EvalNet is trained to assign greater weights to actual training data as opposed to artifical data, and higher weights to artificial data instead of noisy data. As a result, EvalNet secures data augmentation while maintaining NMT performance. EvalNet utilizes three characteristics to assess the quality of parallel data. The loss value is the first of them. The second is the similarity in meaning between the source sentence and the target sentence. As for the third is the cross-attention map that exists between the encoder-decoder components of the Transformer model. The cross-attention map in MT obliquely denotes the relationship between a source and a target sentence. So, noisy parallel sentences might not have the same cross-attention as normal ones. Fully connected layers transform semantic similarity and loss value into feature vectors. LSTM layers transform cross-attention mappings into feature vectors. The semantic similarity, cross-attention map, and loss value are the three inputs used by EvalNet, while the output is an estimate of the data weights. EvalNet provides the evaluation weights that aid in efficiently and effectively training NMT systems from noisy and normal data. Several trials have proven that EvalNet outperforms previous work as a data evaluator. For usage in training NMT systems, artifiacial parallel sentences should have the same meaning regardless of how they were gathered or created. As a result, one of the main characteristics of EvalNet is the semantic similarity of sentence pairings. An embedding vector must first be used to represent a sentence before it can be used to measure the semantic similarity of two sentences. In this work, language-independent BERT sentence embedding is used for embedding a sentence.

In [103], a Transfer Learning Based Semi-Supervised Pseudo Corpus Generation (TLSPG) method is proposed for the translation of zero-resource languages using semi-supervised learning to address zero-shot translation issues and take advantage of similarities among low and zero-resource language pairs. The suggested TLSPG method is based on a hybrid architecture that combines SMT and NMT models. The relationship between language pairings with low and no resources is used by TLSPG to create a pseudo corpus, and TL is used to learn the context of sentences in a semi-supervised manner. As opposed to the multi-lingual ZST scenario where both HRLs and LRLs are considered, the approach here focuses on utilizing a single LRL parallel corpus to develop a MT system for languages with zero available resources. The proposed method consists of three components: Transformer-based semi-supervised learning (TSL), Moses-based semi-supervised learning (MSL), and TL-based creation of pseudo-corpora. The model for zero-resource translation was pre-trained using semi-supervised learning using the TSL and MSL components. The TL-based pseudo-corpus generation component creates a parallel aligned corpus for zero-resource language pairs via pre-trained TSL and MSL modules. Then, after training the MT model using the Transformer or Moses systems, a synthetic parallel corpus is formed by mixing the parallel corpus of the pertinent languages with the pseudo-corpus. TLSPG initially employs the pre-trained TSL or MSL model on monolingual data from the target side of zero-source language pairs. Afterwards, generates monolingual sentences on the source side. The generated source-side monolingual sentences and the target-side monolingual zero-resource language sentences are then parallel aligned in the source-target direction. To produce a synthetic source-target parallel corpus for language pairs with zero resources, TLSPG integrates generated aligned parallel data with a parallel corpus of relevant language pairs. Two models were created specifically for NMT. These are data generated from TSL and data generated from MSL and Transformer models. The mBART model was used to compare the proposed model. In general, the SMT approach gives better results than NMT.

In [104], addresses the problems of domain mismatch in low-resource translation and the lack of low-resource corpora. The Transformer is used as the primary model. Subsequently, the lexical constrained mechanism is applied to the Transformer encoder. In addition, a TL approach is used to overcome corpus limitations. In the pre-processing stage, the authors used an approach called dynamic dictionary. The primary contributions of this research include: a) examining the best data processing strategy to use to enhance neural network performance; b) gathering 60,000 pairs of sentences in English and Vietnamese from the fields of politics,

business, the arts, and sports in order to create a parallel corpus using BT; 3) proposing a new method for low-resource MT through TL based on a lexically constrained model. Token, positional, and segment embedding layers are added to the Transformer model to constrain specific words from references. To investigate the performance of the proposed approach, firstly, the model is trained only on the English-Vietnamese corpus. Second, TL technique are applied through the model to utilize the high-resource language pair. Comparisons with models in different translation directions are made to analyze the decoding speed of the proposed model. It was found that the proposed approach works slightly slower. In addition, the behavior of the model with different beam sizes has been studied, and it has been observed that the results do not improve after the beam size exceeds 30.

In [105], the Dual-lEvel-bAck-tRanslation (DEAR) scheme was proposed. As an extension of NMT, multi-modal NMT uses images or videos as auxiliary information. As a model of NMT, back-translation improves the reducibility of languages. The proposed method is generally a dual-level back translation method using multi-pattern joint learning. It is designed to do back-translation at sentence and concept level. In sentence-level back-translation, the target sentence is accepted as the input of the model to construct the source sentence through a translation model. The model used in the study is the Transformer. Concept level BT is presented in the video under the unique character dynamic visual concept. When a video is given, the first k keyframes are obtained. They are then re-encoded as a new action segment, with the following 32 frames for a keyframe. Thus, action detection and concept labels are obtained. Then, the sentence-based concept attribute is created to synchronize coordination between the sentence and the action. Sentences and action concepts are combined using the joint attention method. For this, a technique called multi-pattern joint learning has been introduced. This method relies on two corpus that share of the Transformer parameters during translation from source-target and target-source. This makes it easier to restrict the input language by combining parameters. Thus, translation at the sentence and concept level is naturally learned jointly. For action capture, the pre-trained Image-Net model was used with fine-tuning on the Kinetics400 dataset. The suggested approach performed better when compared to several techniques in the literature.

In [106], a more advanced embedding method is proposed that allows sharing of the updated results of word embeddings during the optimization of neural networks. The main idea is that the original embedding matrix is replaced by the inner product of two matrices, R and S.Matrix R is the prior information of the relationships between words, which can be acquired through pre-training or self-iterative training. The matrix S, which maintaines the adaptiveness of conventional embedding, is acquired through iterative training within the limitations of the translation. The relation embedding and translation system are initialized and updated at the same time as part of self-iterative training. Pre-training, on the other hand, involves training the relation embedding first, using other systems like LM. By iteratively updating each embedding during the training phase, the new word embedding matrix, on the one hand, contains the relationship between words and is fully mirrored in the entire embedding matrix. The original embedding, which takes up %50 of the size of the Transformer model, is replaced by the relation embedding. The authors state that the proposed method does not lead to any performance loss in most cases and that %85 of the relation embedding elements equal to 0 can be safely removed, thus reducing the model parameters by at least %40. The method proposed in the study was tested on the Transformer model in some scenarios. A Transformer-XL based language model and a BERT model are used for traditional word embedding pre-training and relation embedding training. Traditional embedding, relation embedding, and shared embeddings were pre-trained in Malagasy, Czech, Spanish, Russian, Lithuanian, and English for low-resource translation tasks. The LM was used to pre-train the first four languages independently, then the LM and BERT model was used for the final two. The data usage is greatly improved in the study, and even though the training data is minimal, the method is able to capture the key features of the language better and thus achieve higher performance improvements. The proposed model outperformed the Transformer model in all translation tasks with fewer parameters.

In [107], an effective method for improving NMT performance in low-resource languages and utilizing monolingual data is proposed using the Wolaytta-English pair. Two primary objectives of the study: a) training a model on the existing Wolaytta-English parallel corpora (base model) and self-learning; b) training the base model on a combination of the original and synthetic corpus using a fine-tuning approach. TThe following questions are addressed in this study: Does the performance of NMT for a low-resource language pair enhance by using only single-language data on the source side? Does the performance of NMT when using English as the source language enhance when using monolingual data from a low-resource language? Three main experiments were carried out in the study. LSTM encoder-decoder, bi-LSTM encoder-decoder, and Transformer were used for basic experiments. The Transformer model, which performed the best among the three main models, was chosen to build a artificial English corpus using the Wolaytta monolingual data. A self-learning technique was applied to Transformer model by merging the pseudo-parallel training data with the original parallel data. To create the final NMT model, TL was used to fine-tuned the self-trained NMT model using the Wolaytta-English data. Self-training NMT model with both in-domain and mixed validation sets were used during the fine-tuning. When combining artificial and original parallel data for training and using original corpora for validation, and testing, using only source-side monolingual data was found to improve the

success of NMT in both translation directions for Wolaytta as a low-resource language. Using the original parallel corpora to fine-tune NMT models that were trained on both artificial and original data has demonstrated enhanced NMT performance in both translation directions for the pair of Wolaytta-English.

In [108], an electrical engineering corpus was used for model training, and the issues with the MT model losing the core information as well as varying emphasis on multi-layer information were looked into. Various methods have been used to fused the output vectors of every layer in the encoder. A vector fusion-based multi-attention mechanism translation model is developed on the basis of this, and the decoder component is improved. Thus, the enhanced model gains a more thorough domain knowledge of the source language at the encoder side and enables this knowledge to be better exploited in the decoding phase to enhance the translation success of the model. This study uses Transformer as the basic model. In the multi-layered structure of the encoder units, each layer has different contributions regards to syntax and lexical information in the output vector. When layers are repeatedly stacked on top of one another, the output vector closest to the upper layers concentrates more on grammar, and the output vector of the unit closest to the lower layer concentrates more on the lexical meaning of the source language. As a result, various vector fusion techniques are employed in this work to combine the output vectors from various encoder units before passing them to the decoder. This leads to an enhanced source language representation, consequently improving the translation efficiency of the model. Four techniques-average, additive, weight, and gated fusion-were used in the system fusion experiment and the encoder internal vector fusion experiment, both of which used Transformer as the basic model. Among the fusion methods, the weight-fusion method achieved the best results, and subsequently, a vector fusion-based NMT model with different attention mechanisms is suggested.

In [109], explores the creation of superior NMT models for the resource-poor Kazakh language. First, existing methods for expanding data sizes for low-resource languages like forward translation, BT, and TL are explored. The most common seq-to-seq NMT designs, RNN, Bi-RNN, and Transformer, are described in detail, along with their features, characteristics, and schematics. Then, the ways of creating a Kazakh-English parallel corpus and the training methods of NMT models are explained systematically. For this, a large corpus of 308.000 Kazakh-English sentences was created by combining 205.000 monolingual Kazakh sentences from scientific papers translated with the Promt MT system with 175.000 parallel sentences collected from official government online sources. LSTM, Bi-LSTM, and Transformer architectures of the OpenNMT framework are used to train advanced NMT models, and the results are shared. The best results were obtained with the Bi-LSTM encoder-decoder architecture.

In [110], attempted to overcome the issue that current data augmentation techniques cannot be used in both high- and low-resource settings at the same time. The features and constraints of existing data augmentation methods are analyzed, and a data augmentation method for NMT in a scenario-independent approach is proposed. To further improve the training corpus, the approach combines BT with low-frequency word substitution. Substituting uncommon words for more frequent ones increases the variety of the training data, allows the model to learn to translate a wider variety of words and phrases, and enhance translation accuracy. This prevents the model from overfitting the limited training data. The proposed framework uses additional language model, word frequency modification, and syntax error correction modules. The existing limited-size bilingual and a sizeable target monolingual corpus are first used to build a BT model from the target-source direction. In order to make the final generated corpus parallel to the original one, this paper employs word substitution and uses the grammar error correction module to remove grammatical errors. Subsequently, the generated corpus is combined with a bilingual corpus. The WMT2015 English-German corpora is used in the study, and the high- and low-resource settings are compared in two aspects. An experiment that compares several networks is first carried out. Secondly, a comparative experiment is carried out to relevant data augmentation studies. The proposed method is compared with the RNNSearch, ConvS2S, and Transformer models and the way they are used with different data augmentation methods existing in the literature. The Transformer model outperforms the RNNSearch and ConvS2S models regarding overall translation performance, regardless of whether it is a high or low-resource scenario. Regarding overall translation performance, regardless of whether it is a high- or low-resource scenario, the Transformer model surpasses the RNNSearch and ConvS2S models.

In [111], linguistic attributes are used to create a bidirectional NMT system between the Sanskrit-Malayalam languages. To enhance translation efficiency, the text-based MT system uses linguistic features such as morphological features, POS-tags, and word sense disambiguation (WSD). The Transformer based Sanskrit-Malayalam translation model comprises six distinct modules, incorporating linguistic features. In this study, in addition to text data, manually created audio data is also used. The corpus to be used for the NMT system was recorded by vocalization and used in a multimodal structure. The Transformer model comprises a feature extraction block and a multi-mode feature-level fusion module. This model uses a Sanskrit-Malayalam corpus and speech data. Text in Sanskrit and Malayalam is provided as input to the Transformer model. Speech signals are sent to the Wavelet Transform (WT), Sequential Mapped True Transformation (SMRT), and GCB-based True Transformation (GMRT) modules for obtaining features. Various feature-level multimodal fusion techniques are used to merge the features from the WT/SMRT/GMRT module with the content vector

TABLE 6. Details of low-resource NMT studies. This table contains general information about the studies. It was used for RQ1 and RQ3 answers. The language pairs in the table do not refer to translation directions.

Refs.	Area(s) of focus in the study	Language pair(s) being worked on	Corpora used / amount of data used for the models
[68]	—	English - Mizo	English - Mizo blog posts and Mizoram government web pages, created by themselves / English - Mizo: 29.973
[69]	Transfer Learning	Uyghur - Chinese	CWMT / Uyghur - Chinese : 0.35M WMT'16 / English - Turkish : 0.2M Union Corpus / English - Chinese :15M
[70]	Use of Additional Grammar / Morphologic Feature	Korean - Vietnamese	National Institute of Korean Languages Learner Dictionary and some multilingual articles, created by themselves / Korean - Vietnamese: 454.751
[71]	Hybrid Architecture, Ensemble Learning	English - Hindi	HindiEnCorp / English - Hindi: 273.880
[72]	Knowledge Distillation	Uyghur - Chinese Mongolian - Chinese	Bilingual data source unknown / Uyghur - Chinese: 170k Mongolian - Chinese: 260k Chinese monolingual data : CWMT2017 / 110M
[73]	Hybrid Architecture, Use of Additional Grammar / Morphologic Feature, Data Augmentation	Sanskrit - Hindi	Sanskrit News, ILCI and Department of Sanskrit Studies UoH Digital Sanksrit Corpora / Sanskrit - Hindi : 102.760 - 50k Sanskrit monolingual data : Computational Linguistics J / 2.3M
[74]	GAN Structure / Adversarial Training	Uyghur - Chinese	CCMT 2019 New Translation / Uyghur - Chinese : 0.17M
[75]	Multi-lingual Translation, Use of Pre-trained Model	English - Guarani English - Kazakh English - Dutch English - Turkish English - Japanese English - Myanmar English - Vietnamese English - Arabic English - Italian English - Korean English - Nepali English - Romanian English - Sinhala	WMT'19 / English - Guarani: 10k English - Kazakh: 91k IWSLT'15 / English - Vietnamese: 133k WMT'17 / English - Turkish:207k IWSLT'17 / English - Japanese: 223k English - Korean: 230k English - Dutch: 237k English - Arabic: 250k English - Italian: 250k WAT'19 / English - Myanmar: 259k FLoRes / English - Nepali: 564k English - Sinhala: 647k WMT'16 / English - Romanian: 608k

TABLE 6. *(Continued.)* **Details of low-resource NMT studies. This table contains general information about the studies. It was used for RQ1 and RQ3 answers. The language pairs in the table do not refer to translation directions.**

[76]	Hybrid Architecture, Data Augmentation	English - Hindi Hindi - Bengali Old Eng. - New Eng.	ILCI (health, tourism) / English - Hindi (health): 23k English - Hindi (tourism): 23k Hindi- Bengali (health): 22.012 Hindi- Bengali (tourism): 21.950 IIT Bombay (judical) / English - Hindi: 5561 The Homilies of the Anglo-Saxon Church and its modern translate / Old Eng. - New Eng.: 2674
[77]	Data Augmentation	Chinese - Vietnamese	Created from web pages / Chinese - Vietnamese: 50k Chinese monolingual data: WMT'19 / data size unkown
[78]	Use of Additional Grammar / Morphologic Feature, Corpora Creation	English - Korean Korean - Vietnamese	Created by themselves / English - Korean: 969.194 Korean - Vietnamese: 412.317
[79]	Data Augmentation	English - Twi English - Xitsonga English - Afrikaans Japanese - Russian English - Setswana	English - Twi paralel bible corpus / English - Twi: 122.400 JW300 / English - Afrikaans: 995.740 English - Xitsonga: 766.752 English - Setswana: 791.068 OPUS / Russian - Japanese: 18.462 OPUS Monolingual data: English: 186.537 Twi: 91.642 Russian: 75.402 Japanese: 165.742
[80]	Use of Pre-trained Model, Transfer Learning, Attention Mechanishm	English - French English - Chinese English - German English - Spanish	IWSLT'14 / English - German: 160k English - Spanish: 183k English - French: 236k English - Chinese: 235k
[81]	Data Augmentation, GAN Structure	English - Uyghur English - Vietnamese English - Azerbaijani English - Uzbek English - Turkish English - German English - Hindi	Tanzil / English - Azerbaijani: 21.2k English - Hindi: 182k English - Uyghur: 81.1k English - Uzbek: 134.6k English - Turkish: 141.9k IWSLT'14 / English - German: 160k IWSLT'15 English - Vietnamese: 140.5k
[82]	Use of Pre-trained Model, Attention Mechanishm, Unsupervised Learning	Korean - Japanese English - Korean	AIHUB / English - Korean: 50k Data source unknown / Korean - Japanese: 50k

TABLE 6. *(Continued.)* **Details of low-resource NMT studies. This table contains general information about the studies. It was used for RQ1 and RQ3 answers. The language pairs in the table do not refer to translation directions.**

			Monolingual data: Korean Contemporary Corpus / Koeran: 5M WMT'19 and JParaCrawl / Japanese: 5M WMT'17 and WMT'19 / English: 5M
[83]	Use of Additional Grammar / Morphologic Feature, Use of Pre-trained Model	Vietnamese - Korean	Created by themselves / Vietnamese - Korean: 412.317
[84]	GAN Structure / Adversarial Training	English - Turkish English - Azerbaijani English - Hindi English - Tagalog English - Danish English - Norwegian English - Korean	Tatoeba / English - Turkish: 7k English - Azerbaijani: 2.2k English - Hindi: 2.2k English - Tagalog: 2.2k English - Danish: 7k English - Norwegian: 2.2k English - Korean: 2.2k
[85]	Use of Pre-Trained Model, Model Size / Parameter Reduction	English - Italian English - Spanish English - Dutch	IWSLT'14 / Unknown data size
[86]	Attention Mechanism, Vector Fusion, Model Size / Parameter Reduction	English - German English - Turkish English - Vietnamese	News Commentary v11 (NC11) / English - German: 226.822 IWSLT'14 / English - German: 160.239 WMT'18 / English - Turkish: 207.373 IWSLT'15 / English - Vietnamese: 133.314
[87]	Reinforcement Learning, Domain Adaptation, Transfer Learning	Hindi - Nepali Hindi - Marathi	TDIL / Hindi - Nepali (Agriculture): 12k Hindi - Nepali (Entartainment): 22k WMT'20 Hindi - Marathi (News): 10.349 Hindi - Marathi (PMIndia): 23.897
[88]	Use of Pre-trained Model, Data Augmentation	Chinese - Vietnamese Japanese - Vietnamese	TED Talk / Chinese - Vietnamese: 244.076 Japanese - Vietnamese: 244.417 ALT / Chinese - Vietnamese: 18.088 Japanese - Vietnamese: 18.088 Chinese monolingual data: TED Talk / 244.076 CCAlinger / 2.5M Japanese monolingual data: TED Talk / 244.417 CCAlinger / 2.5M
[89]	Embedding, Multi-lingual Translation, Zero-shot Translation	English - Hindi English - Malayalam English - Assamese English - Telugu English - Tamil English - Manipuri English - Bengali	PMIndia / English - Assamese: 9.372 English - Bengali: 56.866 English - Hindi: 83.258 English - Manipuri: 12.321 English - Tamil: 63.403 English - Telugu: 66.155 English - Malayalam: 57891

TABLE 6. *(Continued.)* **Details of low-resource NMT studies. This table contains general information about the studies. It was used for RQ1 and RQ3 answers. The language pairs in the table do not refer to translation directions.**

			PIB-v1.3 / English - Bengali: 93.560 English - Hindi: 269.594 English - Tamil: 118.759 English - Telugu: 44.888 English - Malayalam:44.986 TDIL+vikaspedia / English - Manipuri: 18.056 Monolingual data: PMIndia/ English: 148.151 Assamese: 46.378 Bengali: 147.288 Hindi: 187.858 Manipuri: 68502 Tamil: 122.410 Telugu: 131.865 Malayalam: 148.151
[90]	Reinforcement Learning, GAN Structure	Hindi - Nepali Hindi - Gujarati Hindi - Maithili Hindi - Punjabi Hindi - Urdu	CVIT-PIB / Hindi - Gujarati: 15k WMT'19, OPUS and TDIL / Hindi - Nepali: 133k OPUS / Hindi - Punjabi: 200k Hindi - Maithili: 93k Hindi - Urdu: 100k
[91]	Vector Fusion	English - German	IWLST'14 / English - German: 16k
[92]	Transfer Learning	Arabic Dialects	MADAR / Arabic Dialects: 40k PADIC / Algerian Arabic Dialect: 10k
[93]	Transfer Learning	Chinese - Azerbaijani Chinese - Uzbek	Tanzil / Chinese - Azerbaijani: 20.1k Chinese - Uzbek: 10.5k Chinese-LDC / Chinese - Uyghur: 10.9k OpenSubtitle 2016 / Chinese - Arabic: 5.1M Chinese - Farsi: 1.4M Chinese - Turkish: 4.4M
[94]	—	English - Punjabi	Created by themselves / English - Punjabi: 259.623
[95]	Use of Additional Grammar / Morphologic Feature	English - Thai Thai - Myanmar English - Myanmar	ASEAN-MT / English - Thai: 20k Thai - Myanmar: 20k English - Myanmar: 20k
[96]	Use of Pre-trained Model, Vector Fusion, Copora Creation	English - Turkish	Abstract of scientific theses, created by themselves / English - Turkish: 1.217.300
[97]	Multi-lingual Translation, Attention Mechanishm, Vector Fusion	English - Spanish English - Portuguese English - Chinese English - Dutch English - Russian English - Romanian	IWSLT'14 / English - Spanish: 180.850 English - Dutch: 167.943 English - Portuguese: 171.903 English - Romanian: 182.141 English - Russian: 178.165 English - Chinese: 179.901

TABLE 6. *(Continued.)* **Details of low-resource NMT studies. This table contains general information about the studies. It was used for RQ1 and RQ3 answers. The language pairs in the table do not refer to translation directions.**

[98]	Embedding, Adversarial Training	English - Chinese	United Nations parallel corpus / English - Chinese: 90.278
[99]	Semi-supervised Learning, Data Augmentation	English - Manipuri	vikaspedia, TDIL and PMIndia / English - Manipuri: 25.475
[100]	Use of Additional Grammar Feature / Morphologic Information	English - Slovene	EuroParl / English - Slovene: 618.516
[101]	Embedding	English - French English - Czech English - Basque	IWSLT'16 / English - French: 219.777 English - Czech: 11.242 English - Basque: 89.413
[102]	Data Augmentation, Use of Pre-trained Model	English - Korean English - Myanmar	IWLST'17 / English - Korean: 234.080 AIHUB / English - Korean: 234.080 WAT / English - Myanmar: 256.100
[103]	Transfer Learning, Zero-shot Translation, Data Augmentation, Semi-supervised Learning	Hindi - Bhojpuri Hindi - Magahi	OPUS and TDIL / Hindi - Nepali: 136.991 Monolingual data: LoResMT2020 / Hindi: 473.605 Bhojpuri: 91.131 Magahi: 148.606
[104]	Embedding, Transfer Learning	English - Vietnamese	Created by themselves / English - Vietnamese: 60k
[105]	Multi-modal Translation, Data Augmentation	English - Chinese English - Turkish	VATEX / English - Chinese: 108.085 and 21.617 video MSVD-Turkish / English - Turkish: 46.635 and 1200 video
[106]	Embedding, Modal Size / Parameter Reduction, Use of Pre-trained Model	Russian - Malagasy Spanish - Czech Spanish - Russian English - Lithuanian	GlobalVoices v2018q4 / Russian - Malagasy: 80k Spanish - Czech: 15k Spanish - Russian: 170k WMT'19 / English - Lithuanian: data size unknown
[107]	Transfer Learning, Data Augmenatation	English - Wolaytta	English - Wolaytta corpus taken from a different study Monolingual data: Created by themselves / Wolaytta: 40k
[108]	Vector Fusion	English - Chinese	Created by themselves / English - Chinese: 190k
[109]	Data Augmentation, Corpora Creation	English - Kazakh	created the bilingual and monolingual data themselves / English - Kazakh: 205k Kazakh monolingual: 175k

TABLE 6. *(Continued.)* **Details of low-resource NMT studies. This table contains general information about the studies. It was used for RQ1 and RQ3 answers. The language pairs in the table do not refer to translation directions.**

[110]	Data Augmentation	English - German	WMT'15 / English - German: 400k Monolingual data: News Crawl Coprus / German: 2M
[111]	Use of Additional Grammar / Morphologic Feature, Multi-modal Translation, Vector Fusion	Sanskrit - Malayalam	Digital Corpus of Sanskrit, Open Subtitles, Computational Linguistics R&D at JNU-India, MTIL and TDIL / Sanskrit - Malayalam: 254.700
[112]	Use of Pre-trained Model	English - Assamese	EnAsCorp 1.0 / English - Assamese: 203.315 Samanatar / English - Assamese: 138.353 Monolingual data: corpus taken from a different study / Assamese: 2.764.181 English: 3.341.688

produced by the encoder network. The decoder employs this altered context vector to generate the correct translation. Methods called max/min/average fusion are used to conduct fusion. The WT, SMRT&GMRT and average fusion approach produced the greatest outcomes in both translation directions.

In [112], to enhance the English-Assamese NMT system, the potential benefit of pre-alignment and pre-trained LMs is studied. In this work, guided alignment is used together with the Transformer model, and parallel corpora of EnAsCorp1.0 and Samanantar are utilized to improve translation success in both directions. The FastAlign tool and the idea of guided alignment in Transformer based NMT were utilized to obtain token alignment knowledge from source-target parallel sentences. In addition, the alignment information obtained by FastAlign and SimAlign is implemented in Transformer-based NMT. It was found that the SimAlign technique and the Transformer-based NMT provided better translation in both directions than the FastAlign technique. It has been observed that there is an enhancement in both translation directions with the pre-trained language model. In addition, English-Spanish and English-Bengali language pairs were used for comparative analysis using pre-alignment. When the findings were examined, higher translation accuracy was obtained with the SimAlign technique based on pre-trained multi-lingual contextual embeddings, with or without previous alignment information based on FastAlign. In addition, when the pre-trained LM is used, translation accuracy is even higher for longer sentences.

B. RQ1: WHAT IS THE FOCUS OF WORK IN LOW-RESOURCE NMT AND ON WHICH LANGUAGE PAIRS ARE STUDIES CONDUCTED?

With this research question, it is aimed to examine which applications/directions are emphasized in the field of low-resource NMT and which language pairs are used and how often.

Investigating the 45 studies selected for review, it is seen that different applications are made to increase the success of translation in the field of low-resource NMT. In order to classify the studies examined in certain aspects, the focus areas for each study were determined to be a maximum of 4 keywords. Although some studies have achieved successful results for the language pair studied, unlike other studies, they do not focus on any point. In other words, these studies only apply NMT systems that already exist in the literature on one or more language pairs. Due to this situation, such studies do not have a specific focus on any application. In Table 6, the direction of the studies and the low-resource pairs used in the studies or other language pairs used in the low-resource setting are shared.

In Table 7, information about the number of studies in the focused areas of the studies examined is shared. When Table 7 is examined, it is seen that a total of 20 different aspects are focused on for low-resource NMT. The most commonly used aspects are data augmentation with 13 studies, the use of pre-trained models with 10 studies, and the use of transfer learning with 9 studies.

When looking at the most commonly used methods, it is seen that they generally try to cope with the lack of corpora

TABLE 7. Focused directions in the field of low-resoruce NMT and the number of studies using these directions.

Low-Resource NMT (NLP) Directions	Research Papers
Data Augmentation	13
Use of Pre-trained Model	10
Transfer Learning	9
Use of Additional Grammar / Morpohologic Fueature	7
Vector Fusion	6
Embedding	5
GAN Structure / Adversarial Training	4
Attention Mechanism	4
Coprora Creation	4
Hybrid Architecture	3
Model Size / Parameter Reduction	3
Multi-lingual Translation	3
Reinforcement Learning	2
Zero-shot Translation	2
Semi-Supervised Learning	2
Multi-modal Translation	2
Ensemble Learning	1
Knowledge Distillation	1
Unsupervised Learning	1
Domain Adaptation	1

in the low-resource NMT field. Data augmentation includes methods used to increase corpus size such as back-translation, forward-translation, etc. When looking at pre-trained models, using the BERT model comes to the fore. It is seen from the studies examined that the use of models such as BERT with a language-specific pre-training provides higher translation performance. The transfer learning method, on the other hand, is a method that is frequently used in the field of machine learning in general and is frequently used here to cope with the lack of corpus. It is often preferred to train a model with a high-resource language pair and then apply this model to a low-resource language pair. In the field of low-resource NMT, it is seen in some studies that it is aimed to increase the success of translation by including various grammatical features and morphological information in the models. However, the use of these features varies according to the extent to which the languages to be studied have the tools to provide these features. In addition, the inclusion of these features in the models may cause additional costs on the models if there are no ready tools for the language pairs studied. Looking at some of the reviewed studies, it is seen that changes have been made in the embedding and attention layers of NMT models. Various techniques are employed to fuse output vectors from specific units within the encoder or decoder. This is done to tackle the issues of information loss in the translation model and varying levels of emphasis on multi-layer information. Models built using these methods can improve translation by combining information from different layers. Since embeddings contain direct representations of words or sentences, changes made in this section can achieve successful results if they provide better representations. It is seen that the changes made in the attention mechanishm are generally made in order to include the additional features used in the models or to provide a stronger representation. When we look at the reviewed studies, it is seen that there are five studies using unsupervised approaches. GAN architecture and the adversarial training approach used in these structures have recently been used in the field of NMT, and it is seen that there are four studies using this method among the studies examined. In unsupervised architectures, back and forward translation models are usually created first and then these models are used jointly. While the GAN architecture is the basis for NMT studies, the model is based on a generator and discriminator. Reconstruction loss occurs when noisy translations are reconstructed in forward and backward directions, and discrimination loss occurs when the translated text is attempted to be separated from the original text. Using this method can provide a higher-quality translation for LRLs that is more fluid. These methods are used by modifying the adversarial framework by including additional adversarial steps or extra loss functions in the optimization step. Among the reviewed studies, the one that has an unsupervised architecture and does not use a GAN structure is a study on the use of the pre-trained MASS model in the NMT model.

A single model for translation across several languages using accessible linguistic resources in numerous languages in multi-lingual NMT approaches, which deal with translation between multiple language pairs [26]. With this approach, knowledge from multiple languages can be learned collectively and applied to help low-resource languages. Multi-lingual NMT techniques use data from various languages to build models. Information from high-resource languages can be utilized to improve the success on low-resource languages because such systems attempt to represent many languages in the same vector space [43]. When the reviewed studies are examined, it is seen that multi-lingual translation is used in only three studies. The methods used in these studies are multi-lingual pre-training, multi-lingual embedding, and multi-lingual implementation of the attention mechanism. Since input from many languages may be employed at once, multi-lingual approaches are generally significantly more effective than training models on language pairs separately. In addition, one of the most essential features of multi-lingual models is that the model can produce a translation for a language not included in the training data, enabling zero-shot translation. Finally, in multi-modal structures, it focuses on how to utilize non-text data to improve translation quality.

Information about the language pairs used in the studies examined and how many studies they were used in are given in Table 8. A total of 72 different language pairs were used in these studies. When the language pairs used in the studies are examined, it is seen that English language is used to a great extent together with a low-resource language and a total of 49 different uses are made in this way. This is more than half the number of language pairs used, and it results that studies generally focus on the English

TABLE 8. Language pairs used in the studies and their usage numbers (The language pairs in the table do not refer to translation direction).

Language pair used	Number of studies used
English - Turkish	6
English - German, English - Hindi, English - Chinese	5
English - Vietnamese	4
English - Spanish, English - Korean, English - Dutch, English - Myanmar, Chinese - Uyghur	3
English - Kazakh, English - Italian, English - Romanian, English - French, English - Azerbaijani, English - Assamese, English - Korean, English - Manipuri, Chinese - Vietnamese, Japanese - Vietnamese, Korean - Vietnamese, Hindi - Nepali	2
English - Mizo, English - Guarani, English - Japanese, English - Arabic, English - Nepali, English - Sinhala, English - Bengali, English - Thai English - Russian, English - Slovene English - Basque, English - Portuguese English - Czech, English - Tamil English - Lithuanian, English - Wolaytta English - Twi, English - Afrikaans English - Xitsonga, English - Setswana English - Uyghur, English - Uzbek English - Tagalog, English - Punjabi English - Danish, English - Norwegian English - Telugu, English - Malayalam Mongolian - Chinese, Hindi - Sanskrit Hindi - Gujarati, Hindi - Maithili Hindi - Marathi, Hindi - Urdu Hindi - Bhojpuri, Hindi - Magahi Hindi - Bengali, Hindi - Punjabi Chinese - Azerbaijani, Chinese - Uzbek Japanese - Russian, Thai - Myanmar Russian - Malagasy, Spanish - Czech Spanish - Russian, Sanskrit - Malayalam Korean - Japanese, Arabic Dialects Old English - New English	1

language in any direction. In addition, it was also observed that most of the translation processes used a high-resource language on one side or the other. The number of cases where languages known to have high resources, such as English, Chinese, French, Italian, Spanish, German, and Arabic, are used in any translation direction is 53 out of 72. After English, the most used language in any direction is Hindi, with 11 different uses, and Chinese, with 6 different uses. WThe studies show that while low-resource languages are used in any direction, high-resource languages are often used in low-resource settings. It seems that the use of low-resource languages in a translation process with each other is less than in other situations, and in this way, there are 15 different translation directions.The most common language pairs used in the studies were English - Turkish with six studies, English - German, English - Hindi, and English - Chinese with five studies each. Turkish and Hindi are known to be low-resource languages, but German and Chinese are high-resource languages and were used in low-resource settings in these studies (<1M data). In addition, Table 9 provides information on the language families of the languages used in the studies analyzed. Based on this information, it can be said that in the field of low-resource NMT, studies are generally conducted on languages in the Indo-Endonesian language family. Information about which language families the languages belong to is taken from the Glottolog[1] website.

TABLE 9. Language families to which the languages used in the reviewed studies belong.

Language Family	Languages Used
Indo - European	English, German, Hindi, Spanish, Dutch, İtalian, Romanian, French, Assamese, Nepali, Sinhala, Bengali, Russian, Slovene, Portuguese, Czech, Lithuanian, Afrikaans, Punjabi, Danish, Norwegian, Sanskrit, Gujarati, Maithili, Marathi, Urdu, Bhojpuri, Magahi, Bengali, Punjabi
Turkic	Turkish, Uyghur, Kazakh, Azerbaijani, Uzbek
Sino - Tibetian	Chinese, Myanmar, Manipuri, Mizo
Dravidian	Tamil, Telugu, Malayalam
Atlantic - Congo	Twi, Xitsonga, Setswana
Austronesian	Tagalog, Malagasy
Austroasiatic	Vietnamese
Koreanic	Korean
Japonic	Japanese
Tupian	Guarani
Tai - Kadai	Thai
Ta - Ne - Omotic	Wolaytta
Mongolic - Khitan	Mongolian
Afro - Asiatic	Arabic
Isolated Language	Basque

C. RQ2: WHICH DEEP LEARNING METHODS ARE PREFERRED IN LOW-RESOURCE NMT AND WHICH EVALUATION CRITERIA ARE USED?

This research question investigated which methods are most preferred for model building in the field of low-resource NMT and which evaluation criteria are used to assess the results of the models built with these methods.

Table 10 provides information about the deep learning methods used in the rewieved studies and the studies in which they were used. It is seen that all of the methods used to create NMT models in the reviewed studies were created within the encoder-decoder framework. The most commonly used method among the models is the Transformer model, which has been used in 35 studies. The attention mechanism, known to increase translation success for pre-Transformer encoder-decoder frameworks, was used in 17 studies. In these studies, Luong style attention method was used in 10 studies, and Bahdanau style attention method was used in 7 studies. Apart from these, one of the methods used in the models created is the Transformer-based BERT model, which was used in 5 different studies. In addition, it is seen that the

[1] https://glottolog.org/

TABLE 10. The methods used in the reviewed studies and the studies in which these methods were used.

Method used in the study	Studies used
Transformer	[69], [72], [74], [75], [76], [79], [80], [81], [82], [83], [86], [87], [88], [90], [91], [93], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112]
LSTM/BiLSTM + Luong Attention	[68], [71], [77], [78], [90], [92], [94], [99], [100]
BERT	[80], [83], [85], [88], [96]
LSTM/BiLSTM + Bahdanau Attention	[70], [74], [87], [101]
CNN	[84], [90], [105], [110]
GRU/BiGRU + Bahdanau Attention	[76], [84], [110]
GRU/BiGRU + Luong Attention	[73]
LSTM/BiLSTM	[109]

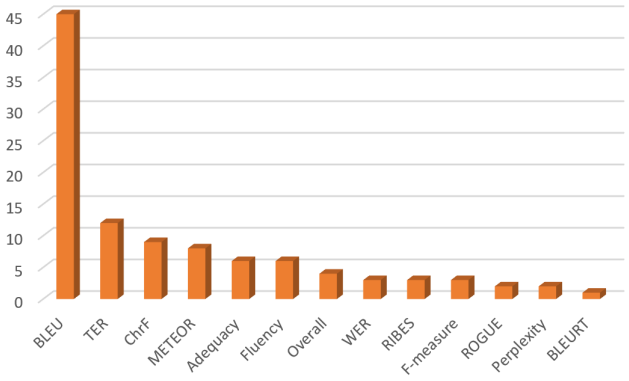


FIGURE 6. Metrics used for the evaluation of models in the studies and the number of uses.

CNN structure is not used much in the field of NMT and has been used in a total of 4 studies. The use of CNN structure is generally preferred when GAN-like structures are applied in the unsupervised NMT domain. The number of studies that do not use any attention mechanism is 1. It is seen that LSTM and GRU structures are generally preferred in encoder-decoder models other than the Transformer model. The LSTM method stands out as the most widely used among these methods. In other words, if the Transformer structure is not used in an NMT model, LSTM is generally preferred. As can be seen in Table 10, since it was proposed in 2017, the Transformer model has been getting the best results in NMT studies and has been used as the mainstream NMT method. In addition, some studies use more than one model and provide comparisons between them.

Machine translation studies proposed in the literature are evaluated with some metrics to assess translation quality and make comparisons. In the reviewed studies, 13 different metrics, namely BLEU, METEOR, TER, WER, RIBES, ChrF, F-measure, ROGUE, Perplexity, Adequacy, Fluency, Overall and BLEURT, were used to evaluate NMT systems. Information about these metrics and their usage is given in Figure 6. Due to the diversity of languages in the world and the variability of languages, there may be more than one translation of a sentence, and which one is correct may vary regionally. There is not yet a standardized approach to evaluate the success of NMT systems [113], [114]. When the reviewed studies are examined, it is seen that 2 types of evaluation criteria are used to evaluate the translation success of the models. These are automatic evaluation and human evaluation metrics. Although the costs of automatic evaluation are lower than human evaluation, the quality of human evaluation is much better. In addition, from the point of view of the studies reviewed, since the proposed models are usually compared with different models, automatic evaluation stands out in terms of speed and provides convenience to researchers. Human evaluation is more costly than automated evaluations in terms of time and human effort. Unlike automatic evaluation, the possibility of inconsistent results should not be ignored since there is a human factor. For these reasons, automatic evaluation criteria are generally preferred in studies. Automatic evaluation is performed by comparing the translations produced by the models with the reference translation, i.e., the correct translation [114]. However, automatic evaluation only captures lexical similarities, and no content or grammar checks are performed. Therefore, sentence structure cannot be properly checked in this way. Table 11 provides information about the evaluation metrics used by the reviewed studies. When the reviewed studies are examined, it is seen that a total of 13 different criteria were used, 10 as automatic evaluation criteria and 3 as human evaluation criteria. Within the scope of this study, the most commonly used automatic evaluation methods BLEU, TER, ChrF, METEOR, and human evaluation criteria, are detailed. In addition to the metrics in Table 11, the BLEURT metric is not included in this table since it is only used in the [104] study.

Since automatic evaluation criteria only consider some aspects of translation quality, the results may be inaccurate. These methods usually use positional information of words to evaluate machine translation. On the other hand, human evaluation assesses machine translation based on adequacy, fluency, and overall rating [68], [113]. Adequacy is the measurement of the amount of meaning of the reference sentence in a machine translation. The fluency metric measures how well the machine translation is generated in the target language without considering the relevance of the machine translation to the reference sentence. The overall rating is the average of the adequacy and fluency values of the machine translation. A translation with high adequacy and fluency values is considered high quality and achieves a high score. All three metrics mentioned above are scored between 1 and 5, with higher values indicating better results.

When the reviewed studies were examined, it was seen that the most commonly used metric was the Bilingual Evaluation Understudy (BLEU) metric, which was found in 45 of the 45 studies. The BLEU score is the most commonly used method for evaluating NMT models. BLEU score, which is an automatic evaluation metric, is utilized to assess how well the translation generated by the MT model resembles the reference translation [115]. Similar to human evaluation,

TABLE 11. Evaluation metrics used in the studies.

	Automatic Evaluation Metrics									Human Evaluation Metrics		
Refs.	BLEU	METEOR	TER	WER	RIBES	ChrF	F-measure	ROGUE	Perplexity	Adequacy	Fluency	Overall
[68]	✓	✓					✓			✓	✓	✓
[69]	✓											
[70]	✓		✓									
[71]	✓											
[72]	✓											
[73]	✓	✓		✓			✓			✓	✓	
[74]	✓											
[75]	✓											
[76]	✓											
[77]	✓	✓				✓						
[78]	✓		✓									
[79]	✓		✓									
[80]	✓											
[81]	✓		✓					✓	✓			
[82]	✓											
[83]	✓	✓	✓									
[84]	✓					✓						
[85]	✓											
[86]	✓											
[87]	✓		✓			✓						
[88]	✓											
[89]	✓									✓	✓	✓
[90]	✓		✓			✓						
[91]	✓											
[92]	✓											
[93]	✓											
[94]	✓		✓							✓	✓	✓
[95]	✓			✓	✓							
[96]	✓	✓	✓						✓			
[97]	✓					✓						
[98]	✓											
[99]	✓					✓				✓	✓	✓
[100]	✓											
[101]	✓											
[102]	✓											
[103]	✓		✓			✓				✓	✓	
[104]	✓	✓						✓				
[105]	✓	✓										
[106]	✓											
[107]	✓		✓			✓						
[108]	✓											
[109]	✓		✓	✓								
[110]	✓											
[111]	✓	✓			✓							
[112]	✓	✓	✓		✓	✓	✓					

the BLEU metric measures the translation's adequacy and fluency. BLEU score is computed through the consideration of three key components: firstly, the precision of n-gram alignment between the machine translation and the reference translation; secondly, the application of a brevity penalty (BP) to counteract potential sentence length bias; and thirdly, the utilization of clipping to appropriately adjust the appearance of continuous words. By dividing the total number of n-grams by the number of matched n-grams, precision is computed. In order to determine the BLEU score, the highest frequency of n-gram matches is counted. The number of n-gram matches is reduced by the maximum number measured in any reference sentence to prevent counting the same n-gram more than once. Short sentences are punished harshly by BLEU since it does not use recall. When the length of the reference sentence is less than the generated sentence, BP is employed to lessen the effect of sentence length on the BLEU score. Since there is almost no human involvement in the evaluation, it is a simple and useful method to assess the quality of the generated translation. However, the BLEU score only uses the n-grams in the sentence, and the results may vary depending on the number of reference sentences. This method only considers word matches, making it difficult to evaluate translations for morphologically rich languages.

The second most commonly used evaluation metric in the reviewed studies is Translation Edit Rate (TER). TER is an automatic evaluation metric utilized to assess the precision of a machine translation [116]. This assessment is conducted by contrasting the machine translation with a reference sentence. It is obtained by calculating the minimum number

of edits required to match the generated translation with the reference sentence. Editing operations include replacing, deleting, adding, and shifting. TER is calculated by dividing the total number of edits by the average word count of the reference sentence. It is a frequently used metric but has some shortcomings. TER focuses only on word-level matches and does not utilize the semantic similarity between the machine translation and the reference sentence. This means that grammatically incorrect translations can score high. Even if a translation is semantically correct, a low score may appear when the words in the sentences do not match exactly. Since the TER only looks at word-level matches, it does not measure the fluency of the machine translation.

Another most commonly used metric is the Character n-gram F-score (ChrF) metric. Unlike the BLEU score, ChrF is calculated by measuring character n-gram overlap instead of word n-grams [117]. This method uses the F-score, which combines character-based n-gram precision and n-gram recall values. N-gram precision represents the percentage of matching n-grams between the machine translation and the reference sentence, while n-gram recall represents the percentage of matching n-grams per character between the machine translation and the reference sentence. Using these two values, the F-score is calculated, and the overlap between the machine translation and the reference sentence is calculated per-character basis. Therefore, it gives better results for character-based languages.

Another metric for measuring translation quality is the Metric for Evaluation of Translation with Explicit ORdering (METEOR). This metric is designed to overcome the limitations of the BLEU score. Unlike BLEU, a weighted F-score is calculated using precision and recall values. The method first aligns the machine translation and the reference sentence to find the longest matching set of words. Words that have identical meaning are considered as the same word during this alignment. Precision and recall are calculated based on the quantity of words that match individually. A penalty is then calculated to reduce the impact of short matches and make longer ones more effective. Adjacent matching phrases are constrained to penalize shorter matches and incentivize longer matches. Its consideration of word stems and words with the same meaning gives it an edge over the BLEU score. This allows the METEOR metric to capture semantic similarity better.

Due to their ease of use and speed, word-based assessment criteria have become popular in recent years for evaluating the quality of NMT systems. However, because these techniques are unable to accurately assess the overall meaning and fluency of machine translation, they are unable to evaluate translation quality effectively [113]. The studies that are under consideration demonstrate that the BLEU measure is the most often applied technique to assess the quality of MT systems. The BLEU metric has limitations, though. The total number of reference sentences may affect the translation outcome because this method only uses n-gram precision [115]. Additionally, there are drawbacks to evaluating MT system translation performance just on the basis of precision [113], [118]. The BLEU metric does not take word stems or synonyms into account; only word matches are taken into account. Additionally, it does not accurately reflect the meaning and sentence structure of the translatation result. Despite the fact that the BLEU score is frequently employed, these restrictions have always necessitated the development of other metrics. One of the advantages of METEOR, which is one of the most commonly used metrics in the analyzed studies, considers stems, synonyms, and word inflections gives it an edge over the BLEU score. As a result, the METEOR metric allows for a considerably better capture of the semantic similarity between the machine translation and the reference translation. Since METEOR employs F-score and penalty functions that take recall and precision values as inputs, it also addresses the issue that punishing short translations of the BLEU metric. According to several research in the literature, the METEOR metric produces findings that are significantly closer to those of human translation than the BLEU score [113], [119]. Despite being frequently used, the BLEU metric has trouble capturing the similarity in semantic content between texts. When the reviewed studies are looked at, it becomes clear that methods like TER, WER, ChrF, and METEOR are utilized in place of this method. In [113], tests were conducted to determine the relationship between some automatic evaluation criteria in the literature and human translation using sentences that are semantically equivalent but have different structures and words. In this study, some metrics used to evaluate MT results were examined by performing a correlation analysis. As a result of this analysis, it is reported that the BLEU score has the lowest correlation score. Contrarily, it was claimed that among the word-based metrics, the METEOR metric had the highest correlation score. This is due to the fact that the METEOR metric uses precision and recall values at the same time and takes into account stems of words and synonyms [113], [120]. Additionally, it was noted that among the word-based metrics, the ChrF measure in the analysis had the highest correlation value. This is assumed to be because the ChrF metric places more emphasis on characters than words [113]. According to this data, metrics that address translation quality from many angles, such as METEOR, TER ChrF, human review, etc., should be evaluated alongside the BLEU score if an NMT system is to be evaluated using automatic evaluation criteria. This is because the BLEU score does not capture all aspects of translation quality and therefore the use of additional metrics is important to better understand the quality of the proposed NMT system.

D. RQ3: WHAT ARE THE CORPORA AND DEVELOPMENT TOOLS USED IN THE STUDIES?

This research question aims to provide information about the corpora used in low-resource NMT and the tools used to build deep learning models in this field. Since many language pairs are used in the reviewed studies, it is seen that

TABLE 12. Monolingual corpora and languages used in the studies.

Used Corpus	Languages
PMIndia	Hindi, Bengali, Manipuri, Tamil, Telugu, Malayalam
OPUS	English, Russian, Japanese, Twi
WMT ('17, '19)	English ('17, '19), Japanese ('19), Chinese ('19)
LoResMT2020	Hindi, Bhojpuri, Magahi
News Crawl Corpus	German
TED Talk	Chinese
CCAlinger	Chinese
Computational Linguistics J	Sanskrit
Korean Contemporary Corpus	Korean
JParaCrawl	Japanese

many different corpora are used. Information about which corpus was used for which language pair in which study and, if used, information about monolingual corpora is given in Table 6. When the studies are examined, it is seen that some studies did not provide complete information about the corpus they used. Therefore, the corpora for which information was available are shared in this table. When the corpora used in the studies were examined, it was seen that 47 different corpora were used in total. The most commonly used corpora are IWSLT with 11 corpora from various years, WMT corpora with nine corpora from various years, and TDIL corpus with six times. In addition to these, in 8 studies, the authors created their corpora and did not use any other corpus. Figure 7 shows information about the corpora used and their usage numbers.

In addition to the bilingual corpora used in the studies, some studies use monolingual data due to their methods. Information about the languages used in this way and the monolingual corpora used for these languages is shared in Table 12.

While creating NMT models in the literature, libraries that provide tools to simplify the creation of models are generally used. In the second part of this research question, it was examined which development environments were preferred for the creation of the models in the studies examined. In line with the information shared in the studies and the accessible information is given in Figure 8. When the studies are analyzed, it is seen that the most used libraries are OpenNMT(py-tf) [121] with 12 uses, Fairseq [122] with 10 uses and Tensorflow[2] with 6 uses. In general, PyTorch-based libraries are preferred for NMT models.

E. EVALUATION

This chapter aims to provide information about the analysis and future studies in the field of low-resource NMT after the research questions have been answered. In addition, all the results obtained by the analyzed studies are given in Table 13. The reviewed studies cannot be directly compared as they usually have different corpora, many different languages, and focus on different aspects. Accordingly, taking the RQ1 results and Table 6 into account, an analysis of the studies

[2]https://www.tensorflow.org/

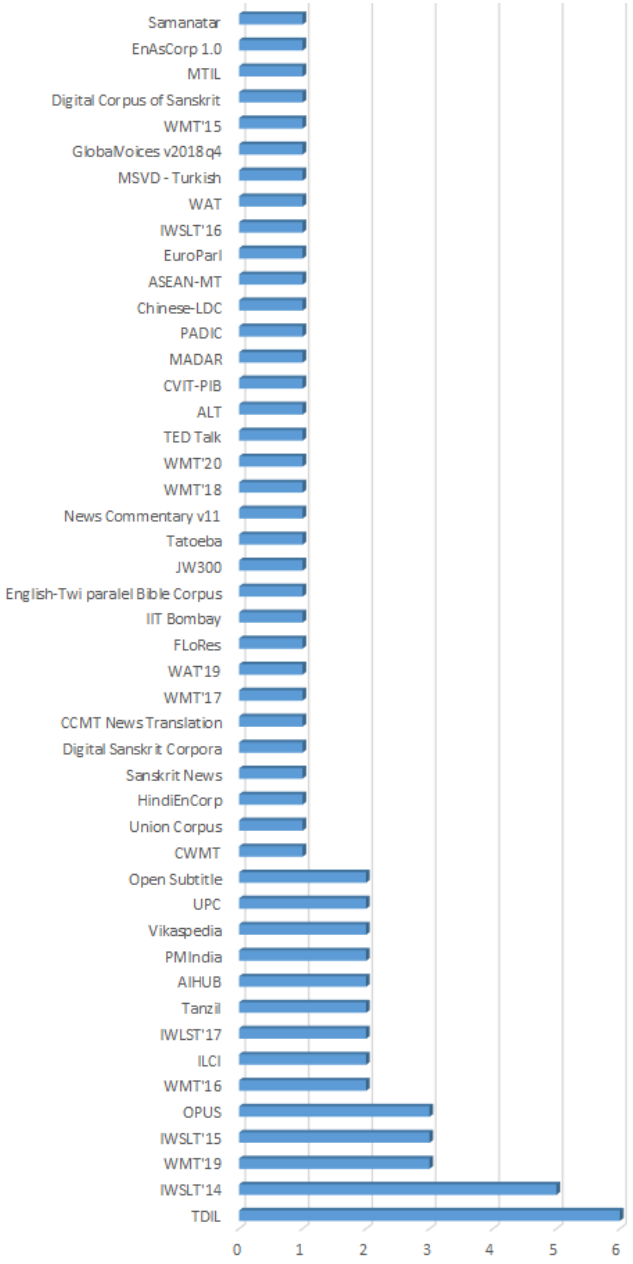


FIGURE 7. Bilingual corpora used in studies.

with the most commonly used language pairs is presented first.

1) ENGLISH-TURKISH (EN-TR)

In the [75] study, the English-Turkish language pair was used in both directions. 17.8 BLEU score was obtained in the En-Tr direction and 22.5 in the Tr-En direction. The results obtained in this study seem relatively low, but the study is multi-lingual. It can be seen as a successful model regarding the method used. In [81], 26.66 BLEU score was obtained in the Tr-En direction. This study is based on the GAN structure, and a data augmentation method

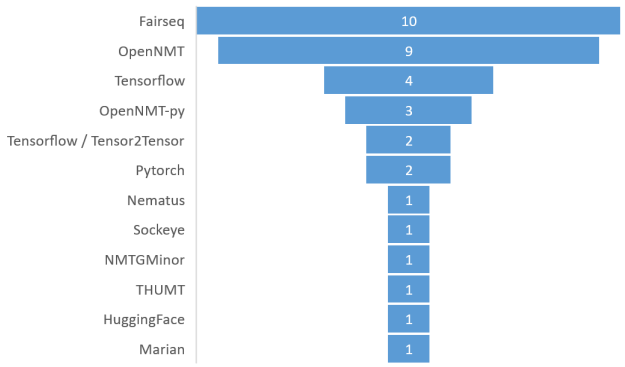


FIGURE 8. Development environments/libraries used in the studies.

is applied. In another study [84], 37.9 BLEU score was obtained in the same translation direction and using the GAN structure. In [86],16.98 BLEU score was obtained in the En-Tr direction. Although this result seems relatively low compared to other studies, the study focused on using the models more efficiently by using fewer parameters. In [96], 45.10 BLEU score was obtained in the Tr-En direction, showing the effect of the pre-trained BERT model on this language pair. The score is high, but this study used a corpus of academic articles. The results on different domains need to be analyzed. Finally, in the [105] study, 44.39 BLEU score was obtained in the Tr-En direction and 36.87 in the En-Tr direction. Unlike other studies, this study is multi-modal in design.

2) ENGLISH-GERMAN (EN-DE)

Reference [80], 31.20 BLEU score in the En-De direction and 38.66 in the De-En direction were obtained. The focus areas of the study are pre-training, transfer learning, and attention mechanism. In [81], 35.14 BLEU score in the De-En direction was obtained using the GAN structure. In [86], 26.12 BLEU score was obtained in the En-De direction with the NC11 corpus. In the De-En direction, 28.46 and 35.79 BLEU scores were obtained using NC11 and IWSLT'14 corpora, respectively. With the knowledge that the test sets used in the studies were not analyzed, it can be said that the IWSLT corpus is more effective in model training. In [91], using the vector fusion method, 35.68 and 35.32 BLEU scores were obtained in the En-De and De-En directions, respectively. In [100], 18.91 BLEU score were obtained in the En-De direction by focusing on data augmentation.

3) ENGLISH-HINDI (EN-HI)

Reference [71] stands out as the only study that uses ensemble learning architecture among the reviewed studies. 19.97 and 21.81 BLEU scores were obtained in En-Hi and Hi-En directions, respectively. In the [76] study, a hybrid structure was used to analyze the models using corpora from different

TABLE 13. Results obtained in the reviewed studies. X-Y: the source language is X, the target language is Y. Only BLEU scores are shared according to RQ2 results.

Ref.	Result(s) obtained (BLEU Score)	Ref.	Result(s) obtained (BLEU Score)	Ref.	Result(s) obtained (BLEU Score)
[68]	Mizo - English : 21.48	[83]	Vietnamese - Korean: 28.22	[98]	English - Chinese: 22.38 Chinese - English: 19.59
[69]	Uyghur - Chinese : 35.51	[84]	Turkish - English: 37.9 Azerbaijani - English: 20.7 Hindi - English: 19.3 Tagalog - English: 22.8 Norwegian - English: 15.3	[99]	English - Manipuri: 7.2 Manipuri - English: 11.9
[70]	Vietnamese - Korean : 25.44 Korean - Vietnamese : 27.79	[85]	English - Italian: 31.81 Italian - English: 34.20 English - Spanish: 37.45 Spanish - English: 42.66 English - Dutch: 32.52 Dutch - English: 38.94	[100]	English - Slovene: 37.37 Slovene - English: 41.80
[71]	English - Hindi : 19.97 Hindi - English : 21.81	[86]	English - German: 26.12 (NC11) German - English: 28.46 (NC11) German - English: 35.79 (IWSLT) English - Turkish: 16.98 English - Vietnamese: 31.89	[101]	English - French: 35.80 Czech - English: 23.75 Basque - English: 21.24
[72]	Uyghur - Chinese : 47.16, with BT 49.01 Mongolian - Chinese : 62.79, with BT 63.02	[87]	Hindi - Nepali Agriculture: 52.50 Entertainment: 35.70 Nepali - Hindi Agriculture: 35.70 Entertainment: 36.91 Hindi - Marathi News: 7.91 PMIndia: 17.85 Marathi - Hindi News: 11.31 PMIndia: 21.54	[102]	English - Korean (IWSLT): 18.8 English - Korean (AIHUB): 27.9 Korean - English (IWSLT): 17.6 Korean - English (AIHUB): 29.4 English - Myanmar: 33.5 Myanmar - English: 23.6
[73]	Sanskrit - Hindi : 61.02	[88]	Chinese - Vietnamese TED Talk: 18.6 (without BERT) ALT: 15.7 Japanese - Vietnamese TED Talk: 23.7 ALT: 14.4	[103]	Hindi - Bhojpuri: 3.78 Hindi - Magahi: 3.18 Bhojpuri - Hindi: 19.49 Magahi - Hindi: 16.14
[74]	Uyghur - Chinese : 32.40	[89]	English - Assamese: 7.6 English - Bengali: 8.5 English - Hindi: 22.1 English - Manipuri: 8.1 English - Tamil: 9.6 English - Telugu: 9.8 English - Malayalam: 6.6 Assamese - English: 18.1 Bengali - English: 21.3 Hindi - English: 31.9 Manipuri - English: 22.6 Tamil - English: 23.9 Telugu - English: 22.7 Malayalam - English: 22.2	[104]	English - Vietnamese: 31.78 Vietnamese - English: 33.52
[75]	English - Guarani: 0.1 English - Kazakh: 2.5 English - Vietnamese: 35.4 English - Turkish: 17.8 English - Japanese: 19.4 English - Korean: 22.6 English - Dutch: 34.8 English - Arabic: 21.6 English - Italian: 34.0 English - Myanmar: 36.9 English - Nepali: 7.4 English - Romanian: 37.7 English - Sinhala: 3.3 Guarani - English: 0.3 Kazakh - English: 7.4 Vietnamese - English: 36.1 Turkish - English: 22.5 Japanese - English: 19.1 Korean - English: 24.6 Dutch - English: 43.3 Arabic - English: 37.6 Italian - English: 39.8 Myanmar - English: 28.3 Nepali - English: 14.5 Romanian - English: 37.8 Sinhala - English: 13.7	[90]	Hindi - Gujarati: 24.8 Hindi - Nepali: 34.2 Hindi - Punjabi: 62.5 Hindi - Maithili: 71.8 Hindi - Urdu: 15.8 Gujarati - Hindi: 27.8 Nepali - Hindi: 32.1 Punjabi - Hindi: 61.5 Maithili - Hindi: 69.6 Urdu - Hindi: 17.1	[105]	English - Chinese: 35.70 Chinese - English: 29.13 English - Turkish: 36.87 Turkish - English: 44.39
[76]	English - Hindi Health: 23.97 Tourism: 19.74 Judical: 25.99 Hindi - English Health: 25.70 Tourism: 24.40 Judical:23.97 Hindi - Bengali Health: 27.04 Tourism: 24.0 Bengali - Hindi Health: 28.90 Tourism: 26.88 Old Eng. - New Eng.: 32.67	[91]	English - German: 35.68 German - English: 35.32	[106]	Russian - Malagasy: 20.26 Spanish - Czech: 4.98 Russian - Spanish: 21.57 Lithuanian - English: 21.73
[77]	Chinese - Vietnamese: 12.38 Vietnamese - Chinese: 6.36	[92]	Arabic Dialects: 35.87	[107]	English - Wolaytta: 16.1 Wolaytta - English: 9.0
[78]	English - Korean: 25.36 Korean - English: 27.45 Korean - Vietnamese: 27.81 Vietnamese - Korean: 25.62	[69]	Azerbaijani - Chinese: 48.62 Uzbek - Chinese: 45.83	[108]	Chinese - English: 37.60
[79]	English - Twi: 19.63 Twi - English: 19.48 English - Afrikaans: 24.09 English - Xitsonga: 19.76 English - Setswana: 21.01 Japanese - Russian: 18.45	[94]	Punjabi - English: 43.23	[109]	Kazakh - English: 49.0
[80]	English - German: 31.20 German - English: 38.66 English - Spanish: 42.3 English - French: 39.8 English - Chinese: 27.9	[95]	English - Thai: 37.03 Thai - English: 35.63 Thai - Myanmar: 25.25 Myanmar - Thai: 25.18 English - Myanmar: 31.73 Myanmar - English: 29.84	[110]	English - German: 18.91
[81]	Azerbaijani - English: 27.59 Hindi - English: 23.68 Uyghur - English: 23.67 Uzbek - English: 21.22 Turkish - English: 26.66 German - English: 35.14 Vietnamese - English: 29.88	[96]	Turkish - English: 45.10	[111]	Samskrit - Malayalam: 43.89 Malayalam - Sanskrit: 42.72
[82]	Korean - Japanese Unsupervised: 29.07 Supervised: 41.96 Japanese - Korean Unsupervised: 32.76 Supervised: 51.69	[97]	English - Spanish: 38.8 English - Dutch: 31.0 English - Portuguese: 39.3 English - Romanian: 27.6 English - Russian: 19.1 English - Chinese: 12.7	[112]	English - Assamese: 15.14 Assamese - English: 19.12

domains separately. The best result in the En-Hi direction is 25.99 BLEU score in the judicial domain, while the best result in the Hi-En direction is 25.70 in the health domain. In [81], a BLEU score of 34.68 in the Hi-En direction was obtained using the GAN structure. In [86], a BLEU score of 19.3 in the Hi-En direction was obtained using the GAN structure. When looking at the results between these two studies, a significant difference is observed. However, the size of the corpus used in [86] is extremely small. In [89], on the other hand, a multi-lingual study is performed, and the BLEU score is 31.9 and 22.1 in the Hi-En and En-Hi directions, respectively.

4) ENGLISH-CHINESE (EN-ZH)

In the [80], 27.9 BLEU score was obtained in the En-Zh direction. In [97], 12.7 BLEU score were obtained in the En-Zh direction in a multi-lingual structure. It should be noted that in this study, different languages are used in the model. In [98], the adversarial training method is applied by focusing on the embedding layer. 22.38 in the En-Zh direction and 19.59 BLEU scores in the Zh-En direction were obtained. In [105], a multi-modal structure is used. 35.70 in the En-Zh direction and 29.13 BLEU scores in the Zh-En direction are obtained. In [108], the authors trained a model on their corpus and obtained 37.60 BLEU score in the Zh-En direction.

Considering the results obtained by the [75], [105] studies, it is seen that in the English-Turkish language pair, relatively better results are obtained when the source language is Turkish. Similarly, higher BLEU scores are obtained in English-German and English-Hindi language pairs when English is the target language. Unlike these language pairs, the English-Chinese pair shows lower results when English is on the target side. In addition to the above analysis, studies with Korean-Vietnamese and Hindi-Nepali language pairs, which are used more than other language pairs but do not have a high-resource language in any direction, were also examined.

5) KOREAN-VIETNAMESE (KO-VI)

Reference [70] examined the use of additional grammatical features and morphological information in these two language pairs. 27.79 BLEU score was obtained for Ko-Vi and 25.44 for Vi-Ko. The same method was followed in [78], and 27.81 in the Ko-Vi direction and 25.62 BLEU scores in the Vi-Ko direction were obtained. In [83], the same corpora was used as in [78], and 28.22 BLEU score was obtained in the Vi-Ko direction. In this study, a pre-trained language model was also used in addition to using additional grammar features. When the two studies are compared, the positive effect of using a pre-trained model on the performance of LRLs can be seen. In addition, it is seen that all studies using this language pair use additional grammar features and morphological information. These features are POS-Tags, WSD, morphological analysis, and word segmentation.

TABLE 14. Future directions in the reviewed studies.

Future Work	Refs.	Freqs.
Use of additional grammar feature / morphologic information	[68], [70], [74], [78], [90], [100], [111]	7
Use of pre-trained models and/or language models	[75], [80], [83], [105], [94], [96], [99]	7
Examination of adversarial learning and GAN structure	[74], [81], [84], [103], [105]	5
Transfer learning	[69], [77], [92], [112]	4
Multi-lingual translation	[73], [94], [98], [112]	4
Data augmentation	[76], [77], [87], [88]	4
Examination of semi-supervised and unsupervised frameworks	[83], [87]	2
Use of reinforcement learning	[87], [90]	2
Zero-shot translation	[87], [89]	2
Creating models with more comprehensive domain space	[89], [96]	2
Use of monolingual data	[99], [107]	2
Multi-modal translation	[111]	1

6) HINDI-NEPALI (HI-NE)

In [87], a study was conducted on domain adoption in agriculture and entertainment domains using reinforcement learning, a rarely used method in NMT. The best result in the Hi-Ne direction is 52.50 BLEU score in agriculture, and the best result in the Ne-Hi direction is 36.91 BLEU score in entertainment. In [90], the reinforcement learning approach was applied by combining it with the GAN structure. 34.2 in the Hi-Ne direction and 32.1 BLEU score in the Ne-Hi direction were obtained.

Finally, future work analysis of the reviewed studies is examined in this section. In this direction, it is aimed to determine the future directions that can be worked on in the field of low-resource NMT. The future work analysis of the reviewed studies is given in Table 14. In this context and line with the reviews, future studies are summarized as follows:

- It is seen in the studies that the use of additional grammar features in NMT models has a beneficial effect on the success of the models. The most commonly used of these grammar features are POS tags, WSD, and morphological analysis. However, these features are generally used in morphologically rich languages. The usage areas of these features can be expanded in future studies. However, it is essential avoid additional costs on the models when using these features in terms of model complexity.
- Using pre-trained models increases the success of NMT models regardless of the data size. Pre-trained models such as BERT, BART, and MASS or language models to be trained on monolingual corpora for low-resource settings will increase success in low-resource scenarios.
- GAN-like models and the adversarial training approach used in these structures have recently succeeded in low-resource NMT. Using these structures in combination with data augmentation and reinforcement learning methods is worth investigating.
- Transfer learning approaches are often preferred for situations where there is insufficient knowledge of

machine learning. Since the lack of corpora is the biggest problem in low-resource NMT, transfer learning approaches similar to hierarchical and pivot-based methods can overcome this shortcoming.

- Better utilization of multi-lingual models in low-resource NMT is a good research topic. Multi-lingual models can be more efficient than bilingual ones as they provide information in multiple languages. In addition, the possibility of zero-shot translation that arises with multi-lingual models is worth investigating, especially for extreme LRLs.
- LLMs work successfully on many NLP tasks and have recently gained a lot of popularity. The effective incorporation of LLMs in low-resource NMT systems, their effectiveness and English-XX translation direction needs to be investigated in more depth.
- In multi-modal NMT, when and how to use different models remains an open problem. A good research topic is finding a suitable method where data such as images, video, or audio are indispensable during translation.
- Domain adaptation has always been a remarkable research topic and has attracted the attention of many researchers. Domain adaptation in NMT is frequently closely tied to parameter fine-tuning, unlike the techniques used in SMT. It is still difficult to solve the issue of unidentified test and out-of-domain translations.
- Especially the models created using the Transformer structure have too many parameters. Working on models that can get competitive results with fewer parameters is a good research topic.
- Automatic evaluation metrics are generally preferred when evaluating studies. However, due to their structure, these methods cannot address all aspects of translation quality. Therefore, human evaluation metrics can be used as a supplement to these metrics. In addition, different metrics that can address all aspects of translation quality can be worked on.

V. CONCLUSION

In this study, an SLR study was carried out to examine the methods used in the field of low-resource NMT. According to the inclusion and exclusion criteria determined in the early stages of the study, 45 studies were selected for review. It was aimed to answer three research questions determined after the relevant studies were determined. The first research question is to identify the areas of focus and language pairs used in the field of low-resource NMT. In the studies, it was seen that studies were carried out in a total of 20 different focus. The most focused study aspects are seen as data augmentation, the use of pre-trained models, and transfer learning. The most studied language pairs in the studies are English-Turkish, English-German, English-Chinese, and English-Hindi. The second research question; it is intended to determine which deep learning methods are used in the low-resource NMT field and which metrics are used to evaluate these methods. It was observed that the Transformer method was mostly used in the models created. Except for the Transformer, Luong attention is mostly used in the LSTM seq-to-seq architecture. It was seen that 13 different metrics were used in total for the evaluation of the studies. The most used metric stands out as the BLEU score. The last research question is; it is about identifying bilingual and monolingual corpora used in studies and preferred development environments. The most used corpora are IWSLT and WMT corpora for various years and TDIL corpus. Finally, when looking at the tools used for model creation, it is seen that the most commonly used tools are OpenNMT and Fairseq. In addition to these, studies were made specifically for the studies in which the most used language pairs were found, and suggestions were made for future studies.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the anonymous reviewers for their invaluable suggestions in putting the present study into its final form.

REFERENCES

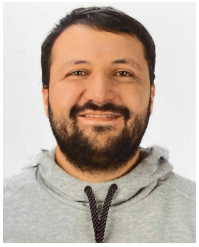
[1] W. Weaver, "Translation," *Machine Translation of Languages*, W. N. Locke and A. D. Booth, Eds. Cambridge, MA, USA: MIT Press, 1955, pp. 15–23.
[2] S. Yang, Y. Wang, and X. Chu, "A survey of deep learning techniques for neural machine translation," 2020, *arXiv:2002.07526*.
[3] Y. Shiwen and B. Xiaojing, "Rule-based machine translation," in *Routledge Encyclopedia of Translation Technology*. Evanston, IL, USA: Routledge, 2014, pp. 186–200.
[4] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguistics*, vol. 19, no. 2, pp. 263–311, Jun. 1993. [Online]. Available: https://aclanthology.org/J93-2003
[5] R. Zens, F. J. Och, and H. Ney, "Phrase-based statistical machine translation," in *Proc. 25th Annu. German Conf. AI, KI*, Aachen, Germany. Berlin, Germany: Springer, Sep. 2002, pp. 18–32.
[6] J. Zhang and C. Zong, "Neural machine translation: Challenges, progress and future," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 2028–2050, Oct. 2020.
[7] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTERSPEECH*, 2010, vol. 2, no. 3, pp. 1045–1048.
[8] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition," in *Proc. INTERSPEECH*, vol. 11, Aug. 2011, pp. 2877–2880.
[9] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2003.
[10] H. Schwenk, D. Déchelotte, and J.-L. Gauvain, "Continuous space language models for statistical machine translation," in *Proc. COLING/ACL Main Conf. Poster Sessions*, 2006, pp. 723–730.
[11] H. Schwenk, "Continuous space translation models for phrase-based statistical machine translation," in *Proc. COLING, Posters*, 2012, pp. 1071–1080.
[12] L. H. Son, A. Allauzen, and F. Yvon, "Continuous space translation models with neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Montreal, QC, Canada, Jun. 2012, pp. 39–48.
[13] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, "Fast and robust neural network joint models for statistical machine translation," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 1370–1380.
[14] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1700–1709.
[15] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[16] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: A case study," 2016, *arXiv:1608.04631*.
[17] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.
[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 3104–3112.
[19] M.-T. Luong, "Neural machine translation," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, 2016.
[20] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. 34th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 1243–1252. [Online]. Available: https://proceedings.mlr.press/v70/gehring17a.html
[21] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," 2019, *arXiv:1901.10430*.
[22] F. Stahlberg, "Neural machine translation: A review," *J. Artif. Intell. Res.*, vol. 69, pp. 343–418, Oct. 2020.
[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[25] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," 2020, *arXiv:2010.12309*.
[26] S. Ranathunga, E.-S.-A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Comput. Surv.*, vol. 55, no. 11, pp. 1–37, Nov. 2023.
[27] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," 2020, *arXiv:2004.09095*.
[28] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," 2015, *arXiv:1511.06709*.
[29] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," 2018, *arXiv:1808.09381*.
[30] V. C. D. Hoang, P. Koehn, G. Haffari, and T. Cohn, "Iterative back-translation for neural machine translation," in *Proc. 2nd Workshop Neural Mach. Transl. Gener.* Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 18–24. [Online]. Available: https://aclanthology.org/W18-2703
[31] A. Poncelas, M. Popovic, D. Shterionov, G. M. de Buy Wenniger, and A. Way, "Combining SMT and NMT back-translated data for efficient NMT," 2019, *arXiv:1909.03750*.
[32] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," 2015, *arXiv:1503.03535*.
[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[34] R. Wang, X. Tan, R. Luo, T. Qin, and T.-Y. Liu, "A survey on low-resource neural machine translation," 2021, *arXiv:2107.04239*.
[35] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 264–280, Dec. 2020.
[36] C. Gulcehre, O. Firat, K. Xu, K. Cho, and Y. Bengio, "On integrating a language model into neural machine translation," *Comput. Speech Lang.*, vol. 45, pp. 137–148, Sep. 2017.
[37] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MASS: Masked sequence to sequence pre-training for language generation," 2019, *arXiv:1905.02450*.
[38] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
[39] L. Wu, Y. Xia, F. Tian, L. Zhao, T. Qin, J. Lai, and T.-Y. Liu, "Adversarial neural machine translation," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 534–549.
[40] P. Koehn and R. Knowles, "Six challenges for neural machine translation," 2017, *arXiv:1706.03872*.
[41] D. Moussallem, M. Wauer, and A.-C.-N. Ngomo, "Machine translation using semantic Web technologies: A survey," *J. Web Semantics*, vol. 51, pp. 1–19, Aug. 2018.
[42] Z. Tan, S. Wang, Z. Yang, G. Chen, X. Huang, M. Sun, and Y. Liu, "Neural machine translation: A review of methods, resources, and tools," *AI Open*, vol. 1, pp. 5–21, Jan. 2020.
[43] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–38, Sep. 2021.
[44] M. S. H. Ameur, F. Meziane, and A. Guessoum, "Arabic machine translation: A survey of the latest trends and challenges," *Comput. Sci. Rev.*, vol. 38, Nov. 2020, Art. no. 100305.
[45] M. Singh, R. Kumar, and I. Chana, "Machine translation systems for Indian languages: Review of modelling techniques, challenges, open issues and future research directions," *Arch. Comput. Methods Eng.*, vol. 28, no. 4, pp. 2165–2193, Jun. 2021.
[46] S. Liu, Y. Sun, and L. Wang, "Recent advances in dialogue machine translation," *Information*, vol. 12, no. 11, p. 484, Nov. 2021.
[47] R. Haque, C.-H. Liu, and A. Way, "Recent advances of low-resource neural machine translation," *Mach. Transl.*, vol. 35, no. 4, pp. 451–474, Dec. 2021.
[48] S. Maruf, F. Saleh, and G. Haffari, "A survey on document-level neural machine translation: Methods and evaluation," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–36, Mar. 2022.
[49] S. A. B. Andrabi, "A review of machine translation for South Asian low resource languages," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 5, pp. 1134–1147, 2021.
[50] S. A. Mohamed, A. A. Elsayed, Y. F. Hassan, and M. A. Abdou, "Neural machine translation: Past, present, and future," *Neural Comput. Appl.*, vol. 33, no. 23, pp. 15919–15931, Dec. 2021.
[51] D. Saunders, "Domain adaptation and multi-domain adaptation for neural machine translation: A survey," 2021, *arXiv:2104.06951*.
[52] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering*, vol. 18, pp. 143–153, Nov. 2021.
[53] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, and A. Birch, "Survey of low-resource machine translation," *Comput. Linguistics*, vol. 48, no. 3, pp. 673–732, Sep. 2022.
[54] Y. Xiao, L. Wu, J. Guo, J. Li, M. Zhang, T. Qin, and T.-Y. Liu, "A survey on non-autoregressive generation for neural machine translation and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11407–11427, Oct. 2023.
[55] F. Li, J. Chen, and X. Zhang, "A survey of non-autoregressive neural machine translation," *Electronics*, vol. 12, no. 13, p. 2980, Jul. 2023.
[56] N. A. Lone, K. J. Giri, and R. Bashir, "Machine translation status of Indian scheduled languages: A survey," *Multimedia Tools Appl.*, early access, pp. 1–29, Apr. 2023.
[57] Y. Zhao, J. Zhang, and C. Zong, "Transformer: A general framework from machine translation to others," *Mach. Intell. Res.*, vol. 20, no. 4, pp. 514–538, Aug. 2023.
[58] R. Aharoni, M. Johnson, and O. Firat, "Massively multilingual neural machine translation," 2019, *arXiv:1903.00089*.
[59] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," 2018, *arXiv:1806.06957*.
[60] O. Firat, B. Sankaran, Y. Al-Onaizan, F. T. Y. Vural, and K. Cho, "Zero-resource translation with multi-lingual neural machine translation," 2016, *arXiv:1606.04164*.
[61] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," 2016, *arXiv:1604.02201*.
[62] R. Dabre, T. Nakagawa, and H. Kazawa, "An empirical study of language relatedness for transfer learning in neural machine translation," in *Proc. 31st Pacific Asia Conf. Lang., Inf. Comput.*, 2017, pp. 282–286.
[63] Y. Cheng, "Joint training for pivot-based neural machine translation," in *Joint Training for Neural Machine Translation*. Singapore: Springer, 2019, pp. 41–54, doi: 10.1007/978-981-32-9748-7_4.
[64] Y. Kim, P. Petrov, P. Petrushkov, S. Khadivi, and H. Ney, "Pivot-based transfer learning for neural machine translation between non-English languages," 2019, *arXiv:1909.09524*.
[65] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2023, *arXiv:2307.03109*.
[66] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Multilingual machine translation with large language models: Empirical results and analysis," 2023, *arXiv:2304.04675*.
[67] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, "How good are GPT models at machine translation? A comprehensive evaluation," 2023, *arXiv:2302.09210*.

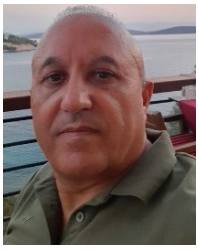
[68] A. Pathak, P. Pakray, and J. Bentham, "English–Mizo machine translation using neural and statistical approaches," *Neural Comput. Appl.*, vol. 31, no. 11, pp. 7615–7631, Nov. 2019.
[69] G. Luo, Y. Yang, Y. Yuan, Z. Chen, and A. Ainiwaer, "Hierarchical transfer learning architecture for low-resource neural machine translation," *IEEE Access*, vol. 7, pp. 154157–154166, 2019.
[70] Q.-P. Nguyen, A.-D. Vo, J.-C. Shin, P. Tran, and C.-Y. Ock, "Korean-Vietnamese neural machine translation system with Korean morphological analysis and word sense disambiguation," *IEEE Access*, vol. 7, pp. 32602–32616, 2019.
[71] D. Banik, A. Ekbal, P. Bhattacharyya, and S. Bhattacharyya, "Assembling translations from multi-engine machine translation outputs," *Appl. Soft Comput.*, vol. 78, pp. 230–239, May 2019.
[72] X. Zhang, X. Li, Y. Yang, and R. Dong, "Improving low-resource neural machine translation with teacher-free knowledge distillation," *IEEE Access*, vol. 8, pp. 206638–206645, 2020.
[73] M. Singh, R. Kumar, and I. Chana, "Improving neural machine translation for low-resource Indian languages using rule-based feature extraction," *Neural Comput. Appl.*, vol. 33, no. 4, pp. 1103–1122, Feb. 2021.
[74] X. Shi, H. Huang, P. Jian, and Y.-K. Tang, "Improving neural machine translation with sentence alignment learning," *Neurocomputing*, vol. 420, pp. 15–26, Jan. 2021.
[75] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020.
[76] S. Sen, M. Hasanuzzaman, A. Ekbal, P. Bhattacharyya, and A. Way, "Neural machine translation of low-resource languages using SMT phrase pair injection," *Natural Lang. Eng.*, vol. 27, no. 3, pp. 271–292, May 2021.
[77] H. Li, J. Sha, and C. Shi, "Revisiting back-translation for low-resource machine translation between Chinese and Vietnamese," *IEEE Access*, vol. 8, pp. 119931–119939, 2020.
[78] V.-H. Vu, Q.-P. Nguyen, J.-C. Shin, and C.-Y. Ock, "UPC: An open word-sense annotated parallel corpora for machine translation study," *Appl. Sci.*, vol. 10, no. 11, p. 3904, Jun. 2020.
[79] M. Adjeisah, G. Liu, D. O. Nyabuga, R. N. Nortey, and J. Song, "Pseudotext injection and advance filtering of low-resource corpus for neural machine translation," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–10, Apr. 2021.
[80] Z. Zhang, S. Wu, D. Jiang, and G. Chen, "BERT-JAM: Maximizing the utilization of BERT for neural machine translation," *Neurocomputing*, vol. 460, pp. 84–94, Oct. 2021.
[81] M. Maimaiti, Y. Liu, H. Luan, and M. Sun, "Data augmentation for low-resource languages NMT guided by constrained sampling," *Int. J. Intell. Syst.*, vol. 37, no. 1, pp. 30–51, Jan. 2022.
[82] Y.-S. Choi, Y.-H. Park, S. Yun, S.-H. Kim, and K.-J. Lee, "Factors behind the effectiveness of an unsupervised neural machine translation system between Korean and Japanese," *Appl. Sci.*, vol. 11, no. 16, p. 7662, Aug. 2021.
[83] V.-H. Vu, Q.-P. Nguyen, E. V. Tunyan, and C.-Y. Ock, "Improving the performance of Vietnamese–Korean neural machine translation with contextual embedding," *Appl. Sci.*, vol. 11, no. 23, p. 11119, Nov. 2021.
[84] M. Sun, H. Wang, M. Pasquine, and I. A. Hameed, "Machine translation in low-resource languages by an adversarial neural network," *Appl. Sci.*, vol. 11, no. 22, p. 10860, Nov. 2021.
[85] J. Guo, Z. Zhang, L. Xu, B. Chen, and E. Chen, "Adaptive adapters: An efficient way to incorporate BERT into neural machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1740–1751, 2021.
[86] L. Gong, Y. Li, J. Guo, Z. Yu, and S. Gao, "Enhancing low-resource neural machine translation with syntax-graph guided self-attention," *Knowl.-Based Syst.*, vol. 246, Jun. 2022, Art. no. 108615.
[87] A. Kumar, A. Pratap, A. K. Singh, and S. Saha, "Addressing domain shift in neural machine translation via reinforcement learning," *Expert Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117039.
[88] T.-V. Ngo, P.-T. Nguyen, V. V. Nguyen, T.-L. Ha, and L.-M. Nguyen, "An efficient method for generating synthetic data for low-resource machine translation: An empirical study of Chinese, Japanese to Vietnamese neural machine translation," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, Art. no. 2101755.
[89] S. M. Singh and T. D. Singh, "An empirical study of low-resource neural machine translation of manipuri in multilingual settings," *Neural Comput. Appl.*, vol. 34, no. 17, pp. 14823–14844, Sep. 2022.
[90] A. Kumar, A. Pratap, and A. K. Singh, "Generative adversarial neural machine translation for phonetic languages via reinforcement learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 190–199, Feb. 2023.
[91] J. Yang, Y. Yin, L. Yang, S. Ma, H. Huang, D. Zhang, F. Wei, and Z. Li, "GTrans: Grouping and fusing transformer layers for neural machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1489–1498, 2023.
[92] A. Slim, A. Melouah, U. Faghihi, and K. Sahib, "Improving neural machine translation for low resource Algerian dialect by transductive transfer learning strategy," *Arabian J. Sci. Eng.*, vol. 47, no. 8, pp. 10411–10418, Aug. 2022.
[93] M. Maimaiti, Y. Liu, H. Luan, and M. Sun, "Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation," *Tsinghua Sci. Technol.*, vol. 27, no. 1, pp. 150–163, Feb. 2022.
[94] K. D. Garg, S. Shekhar, A. Kumar, V. Goyal, B. Sharma, R. Chengoden, and G. Srivastava, "Framework for handling rare word problems in neural machine translation system using multi-word expressions," *Appl. Sci.*, vol. 12, no. 21, p. 11038, Oct. 2022.
[95] Z. Z. Hlaing, Y. K. Thu, T. Supnithi, and P. Netisopakul, "Improving neural machine translation with POS-tag features for low-resource language pairs," *Heliyon*, vol. 8, no. 8, Aug. 2022, Art. no. e10375.
[96] İ. Sel and D. Hanbay, "Fully attentional network for low-resource academic machine translation and post editing," *Appl. Sci.*, vol. 12, no. 22, p. 11456, Nov. 2022.
[97] Q. Wang, J. Zhang, and C. Zong, "Synchronous inference for multilingual neural machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1827–1839, 2022.
[98] Y. Li, Y. Shan, Z. Liu, C. Che, and Z. Zhong, "Transformer fast gradient method with relative positional embedding: A mutual translation model between English and Chinese," *Soft Comput.*, vol. 27, no. 18, pp. 13435–13443, Sep. 2023.
[99] S. M. Singh and T. D. Singh, "Low resource machine translation of English–Manipuri: A semi-supervised approach," *Expert Syst. Appl.*, vol. 209, Dec. 2022, Art. no. 118187.
[100] G. Donaj and M. S. Maučec, "On the use of morpho-syntactic description tags in neural machine translation with small and large training corpora," *Mathematics*, vol. 10, no. 9, p. 1608, May 2022.
[101] I. J. Unanue, E. Z. Borzeshi, and M. Piccardi, "Regressing word and sentence embeddings for low-resource neural machine translation," *IEEE Trans. Artif. Intell.*, vol. 4, no. 3, pp. 450–463, Jun. 2023.
[102] Y.-H. Park, Y.-S. Choi, S. Yun, S.-H. Kim, and K.-J. Lee, "Robust data augmentation for neural machine translation through EVALNET," *Mathematics*, vol. 11, no. 1, p. 123, 2022.
[103] A. Kumar, R. K. Mundotiya, A. Pratap, and A. K. Singh, "TLSPG: Transfer learning-based semi-supervised pseudo-corpus generation approach for zero-shot translation," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 9, pp. 6552–6563, Oct. 2022.
[104] H. Jiang, C. Zhang, Z. Xin, X. Huang, C. Li, and Y. Tai, "Transfer learning based on lexical constraint mechanism in low-resource machine translation," *Comput. Electr. Eng.*, vol. 100, May 2022, Art. no. 107856.
[105] S. Chen, Y. Zeng, D. Cao, and S. Lu, "Video-guided machine translation via dual-level back-translation," *Knowl.-Based Syst.*, vol. 245, Jun. 2022, Art. no. 108598.
[106] Q. Chen, "A smaller and better word embedding for neural machine translation," *IEEE Access*, vol. 11, pp. 40770–40778, 2023.
[107] A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Low-resource neural machine translation improvement using source-side monolingual data," *Appl. Sci.*, vol. 13, no. 2, p. 1201, Jan. 2023.
[108] H. Chen, Y. Chen, and J. Zhang, "Neural machine translation of electrical engineering based on vector fusion," *Appl. Sci.*, vol. 13, no. 4, p. 2325, Feb. 2023.
[109] V. Karyukin, D. Rakhimova, A. Karibayeva, A. Turganbayeva, and A. Turarbek, "The neural machine translation models for the low-resource Kazakh–English language pair," *PeerJ Comput. Sci.*, vol. 9, p. e1224, Feb. 2023.
[110] X. Liu, J. He, M. Liu, Z. Yin, L. Yin, and W. Zheng, "A scenario-generic neural machine translation data augmentation method," *Electronics*, vol. 12, no. 10, p. 2320, May 2023.
[111] C. Rahul, T. Arathi, L. S. Panicker, and R. Gopikakumari, "Morphology & word sense disambiguation embedded multimodal neural machine translation system between Sanskrit and Malayalam," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 105051.

[112] S. R. Laskar, B. Paul, P. Dadure, R. Manna, P. Pakray, and S. Bandyopadhyay, "English–assamese neural machine translation using prior alignment and pre-trained language model," *Comput. Speech Lang.*, vol. 82, Jul. 2023, Art. no. 101524.
[113] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, and H. Lim, "A survey on evaluation metrics for machine translation," *Mathematics*, vol. 11, no. 4, p. 1006, Feb. 2023.
[114] A. Lavie, "Evaluating the output of machine translation systems," in *Proc. 9th Conf. Assoc. Mach. Transl. Amer., Tuts.*, Denver, CO, USA, Oct./Nov. 2010.
[115] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.
[116] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. 7th Conf. Assoc. Mach. Transl. Americas, Tech. Papers*, 2006, pp. 223–231.
[117] M. Popović, "ChrF: Character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop Stat. Mach. Transl.*, 2015, pp. 392–395.
[118] E. Chatzikoumi, "How to evaluate machine translation: A review of automated and human metrics," *Natural Lang. Eng.*, vol. 26, no. 2, pp. 137–161, Mar. 2020.
[119] A. Agarwal and A. Lavie, "METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output," in *Proc. 3rd Workshop Stat. Mach. Transl. (StatMT)*, 2008, pp. 115–118.
[120] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[121] G. Klein, Y. Kim, Y. Deng, V. Nguyen, J. Senellart, and A. M. Rush, "OpenNMT: Neural machine translation toolkit," 2018, *arXiv:1805.11462*.
[122] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (Demonstrations)*. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

BİLGE KAĞAN YAZAR received the bachelor's degree in computer engineering from Ankara University, Ankara, in 2017, and the master's degree in computer engineering from Ondokuz Mayıs University, Samsun, in 2020, where he is currently pursuing the Ph.D. degree in computational sciences. His research interests include machine learning, deep learning, natural language processing, and machine translation.



DURMUŞ ÖZKAN ŞAHİN received the bachelor's degree in computer engineering from Süleyman Demirel University, Isparta, in 2013, and the master's degree in computer engineering and the Ph.D. degree in computational sciences from Ondokuz Mayıs University, Samsun, in 2016 and 2022, respectively. His research interests include machine learning, data mining, text mining, information retrieval, and Android malware analysis.



ERDAL KILIÇ received the bachelor's degree in electrical electronic engineering from Karadeniz Technical University, Trabzon, in 1991, the master's degree in electrical electronic engineering from Karadeniz Technical University, in 1996, and the Ph.D. degree in electrical and electronic engineering from Middle East Technical University, Ankara, in 2005. He is currently a Full Professor with the Department of Computer Engineering, Ondokuz Mayıs University. His research interests include neural networks, machine learning, and data mining.