Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064349

ARTICLE

PNMT: Zero-Resource Machine Translation with Pivot-Based Feature Converter

Lingfang Li[1,2], Weijian Hu[2] and Mingxing Luo[1,*]

[1]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, 611730, China
[2]School of Information Engineer, Inner Mongolia University of Science & Technology, Baotou, 014000, China

*Corresponding Author: Mingxing Luo. Email: mxluo@swjtu.edu.cn

Received: 12 February 2025; Accepted: 27 May 2025; Published: 30 July 2025

ABSTRACT: Neural machine translation (NMT) has been widely applied to high-resource language pairs, but its dependence on large-scale data results in poor performance in low-resource scenarios. In this paper, we propose a transfer-learning-based approach called shared space transfer for zero-resource NMT. Our method leverages a pivot pre-trained language model (PLM) to create a shared representation space, which is used in both auxiliary source→pivot (Ms2p) and pivot→target (Mp2t) translation models. Specifically, we exploit pivot PLM to initialize the Ms2p decoder and Mp2t encoder, while adopting a freezing strategy during the training process. We further propose a feature converter to mitigate representation space deviations by converting the features from the source encoder into the shared representation space. The converter is trained using the synthetic source→target parallel corpus. The final Ms2t model combines the Ms2p encoder, feature converter, and Mp2t decoder. We conduct simulation experiments using English as the pivot language for German→French, German→Czech, and Turkish→Hindi translations. We finally test our method on a real zero-resource language pair, Mongolian→Vietnamese with Chinese as the pivot language. Experiment results show that our method achieves high translation quality, with better Translation Error Rate (TER) and BLEU scores compared with other pivot-based methods. The step-wise pre-training with our feature converter outperforms baseline models in terms of COMET scores.

KEYWORDS: Zero-resource machine translation; pivot pre-trained language model; transfer learning; neural machine translation

1 Introduction

Neural machine translation (NMT) exploits neural networks to automatically produce accurate and fluent translations between languages. Unlike traditional rule-based or statistical machine translation (SMT) approaches [1–3], NMT directly models continuous representations of linguistic units from parallel corpora using neural networks, eliminating reliance on handcrafted linguistic rules and features. This capability has enabled superior performance [4–6], establishing NMT as the dominant approach in machine translation. While end-to-end NMT models for high-resource language pairs have demonstrated impressive results for high-resource language pairs [7,8], their performances heavily depend on the large-scale parallel corpus. Consequently, data scarcity remains a significant challenge for non-English language pairs [9–11]. To address this issue, pivot-based NMT (PNMT) has emerged as an effective solution [12,13], leveraging high-resource "pivot" languages to bridge language pairs with limited or no parallel data, making it suitable for low-resource and even zero-resource scenarios.

PNMT connects source and target languages lacking parallel corpus through a two-step process. As shown in Fig. 1, consider German→French translation using English as the pivot, where X, Y, and Z

represent source, target, and pivot languages, respectively. This process involves pre-training individual NMT models for both source→pivot (Ms2p) and pivot→target (Mp2t) models using large-scale parallel data. Ms2p translates the source sentence into a pivot sentence, which further serves as the input for Mp2t to generate the target sentence. However, pivot models require double decoding operations while the two-step process can lead to the propagation and amplification of translation errors. Moreover, training models are challenging due to the inability of optimization operations to propagate gradients when the decoder generates a pivot hypothesis.

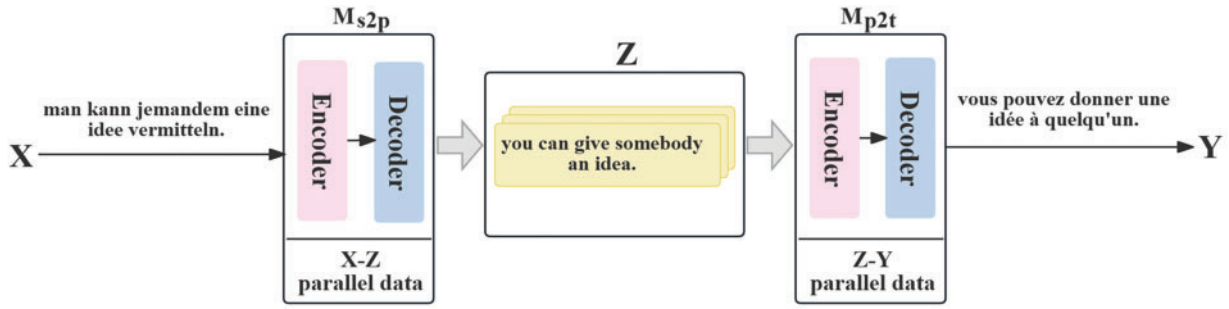

Figure 1: Schematic example of the pivot-based NMT

To resolve these problems, Cheng et al. [14] introduced a joint training approach with shared pivot word embedding to enable interaction between Ms2p and Mp2t. Others maximized the likelihood of the cascading network with a few source-target (s2t) data. The methods underutilize pivot monolingual corpus and then face optimization challenges. Additionally, Tokarchuk et al. [15] replaced the auto-regressive (AR) decoder in Ms2p with a non-autoregressive (NA) decoder trained via reinforcement learning (RL). AR decoder generates target translation word by word, with each step relying on the output of the previous step. While NA decoders generate translations in a single step to improve speed, they often introduce duplicate words or omissions, especially in long sentences, resulting in lower accuracy than AR methods. Alternative strategies involve generating pseudo-parallel source-target data via pivot languages [16,17] and transfer learning techniques [18,19] that adapt knowledge from different but related tasks or domains to improve PNMT. Zoph et al. [20] were the first to utilize parameter transfer from high-resource to low-resource models, and Kim et al. [12] proposed transferring both the Ms2p encoder and Mp2t decoder parameters for zero-shot translation. However, the cascading pivot language method can lead to parameter doubling and inference delay, as well as error accumulation. The transfer-based pivot approaches fail to adequately resolve cross-lingual representation challenges due to the scarcity of parallel has poor performance due to cross-language representation alignment issues.

Recent results explore LLMs for machine translation [21–23], capitalizing on their generative and multilingual capabilities. Prompt engineering [24,25] or in-context learning (ICL) [26,27] can enhance LLM performance for low-resource languages (LRL). Zhu et al. [28] proposed a Language-Aware Neuron Detecting and Routing framework (LANDeRMT) to selectively fine-tune neurons for translation tasks, while Pan et al. [29] built a PrObability-driven Meta-graph Prompter (POMP) to dynamically sample multiple translation paths using auxiliary languages. However, LLMs remain limited for LRL due to their pretraining bias toward high-resource languages.

In this paper, we propose a pivot-based transfer learning [30,31] framework for zero-resource NMT, unifying source, and target language representations via a pivot language space defined by a pre-trained language model (PLM). PLMs learn universal language representations from large monolingual corpora

and can be further fine-tuned for downstream tasks. To align Ms2p and Mp2t with the pivot space, we initialize both models using pivot PLM and employ parameter freezing during pre-training. To mitigate representation space divergence, i.e., the project domain shift [32], we propose a feature converter that maps Ms2p encoder outputs into a space compatible with the Mp2t decoder. Experiments on four language pairs of German→French, German→Czech, Turkish→Hindi (using English as the pivot), and Mongolian→Vietnamese (using Chinese as a pivot) demonstrate that the present method can improve translation performance without parallel data. The main contributions of this work are as follows:

- We proposed a systematic framework that unifies Ms2p and Mp2t representation via PLM-defined pivot space for zero-resource NMT.
- We proposed a feature converter to eliminate the divergence between the source language representation space and the target language representation space.
- We validate the proposed method across four zero-resource language pairs, outperforming existing pivot-based methods in translation quality and efficiency.

2 Background and Motivation

NMT [33,34] is a data-driven approach that utilizes neural networks for translation tasks. In contrast, pivot-based NMT addresses low-resource scenarios by leveraging a third "pivot" language to bridge language pairs with limited parallel data. This approach mitigates the challenge of translating directly between languages with scarce bilingual corpora, instead using pivot as an intermediary to enhance translation quality. In what follows, we briefly introduce concepts of pivot-based NMT.

2.1 Pivot-Based Translation

Pivot-based NMT addresses data scarcity between language pairs using a pivot language. Pivot language is typically chosen based on its substantial parallel corpora with both source and target languages. Define s and t to denote a source sentence and the corresponding target sentence, respectively. Denote an NMT model as $P(t|s;\theta_{s\to t})$, where $\theta_{s\to t}$ denotes one set of parameters. Typically, this model can be trained using parallel corpus $D_{s,t} = \{< s, t >\}$ by maximizing likelihood estimation:

$$\hat{\theta}_{s\to t} = \arg\max_{\theta_{s\to t}} \left\{ \sum_{<s,t>\in D_{s,t}} \log P(t|s;\theta_{s\to t}) \right\} \tag{1}$$

The traditional method often yields poor translation quality when dealing with low-resource language pairs. Intuitively, if there are abundant parallel corpora of source-pivot and pivot-target available for training separate models, one can translate the source sentence to the target sentence through the pivot sentence. Denote p as the intermediate sentence in pivot language. The NMT model for source→target, with pivot language as its connection, can be represented as:

$$P\left(t|s;\theta_{s\to p},\theta_{p\to t}\right) = \sum_{p} P\left(t|p;\theta_{p\to t}\right) P\left(p|s;\theta_{s\to p}\right) \tag{2}$$

where $P\left(t|p;\theta_{p\to t}\right)$ and $P\left(p|s;\theta_{s\to p}\right)$ are probabilities of target and pivot sentence conditional on pivot and source sentence, respectively. This leverages pivot language to facilitate source→target translation. According to high-resource models, the pivot-based methods can be divided into three categories.

Direct Pivot Translation. The procedure contains two steps. First, the source sentence will be translated into a pivot sentence, which is further translated into the target in the second step. It has not trained individual NMT models for translation between source and target languages directly. This method depends on the

reliability of high-quality translation models (Ms2p and Mp2t) but requires twice decoding. Although it increases translation time, this method has good performance in zero-shot language translation.

Data Augmentation Translation. The general data augmentation method is to synthesize parallel data by translating pivot sentences to either source sentences [35] or target sentences [36]. The synthesized data can then be used as a training corpus to train source→target NMT model. The quality and quantity of parallel corpora determine the performance of the translation model.

Model Transfer Training. It involves leveraging the learned knowledge from high-resource models to improve low-resource or zero-resource translation using transfer learning [37,38]. Transfer learning introduced by [20] is initially used in the context of pivot-based NMT. The low-resource model will be initialized with parameters of a high-resource model and then fine-tuned to adapt to the specific task. This can improve the performance of NMT models suffering from parallel data scarcity.

Pivot-based approaches, which address the scarcity of parallel corpora in machine translation, are susceptible to cascaded translation errors. Specifically, errors in Ms2p propagate to Mp2t. This discrepancy arises partly due to significant differences in vocabulary and parameter space between the two models, as parallel data used for training is often loosely correlated or even unrelated.

To enhance pivot-based NMT, one method is to reduce the discrepancy or increase the similarity between Ms2p and Mp2t. Our approach involves sharing the pivot language representation between Ms2p and Mp2t. Additionally, we introduce a feature converter to transform features into forms more familiar to the target decoder. This method aligns representations and features across two models, mitigating the cascaded translation errors and improving overall translation quality in pivot-based NMT.

2.2 Language Representation Space

Most NMT approaches are based on the Encoder-Decoder framework [39,40] and attention mechanism. In this framework, the encoder transforms the input sequence into a continuous representation sequence, often referred to as feature vectors. These vectors play a crucial role, as they satisfy the distributional hypothesis and encapsulate textual data in a meaningful way. The decoder then processes these encoded representations, generating target language output one token at a time. Pivot-based NMT involves using two independently trained sequence-to-sequence models, namely Ms2p and Mp2t. The main goal is to obtain target probability distributions by leveraging an intermediate representation of pivot language, given a source sentence.

We hypothesize that the representation spaces of the source language (German) and pivot language (English) exhibit similarity when conveying the same semantic meaning. To show this, we selected parallel sentences in German and English, encoding their words separately using encoders of the trained Ms2p and Mp2t models. We then extracted the encoded feature vectors and reduced their dimensionality using locally linear embedding (LLE) and principal component analysis (PCA). Fig. 2 illustrates the distributions of source and pivot feature vectors after dimensionality reduction. It indicates that the representation spaces for different languages when expressing identical semantics, demonstrate a measurable degree of similarity. This suggests the feasibility of establishing a connection between two language pairs by exploiting the pivot language's representation space.

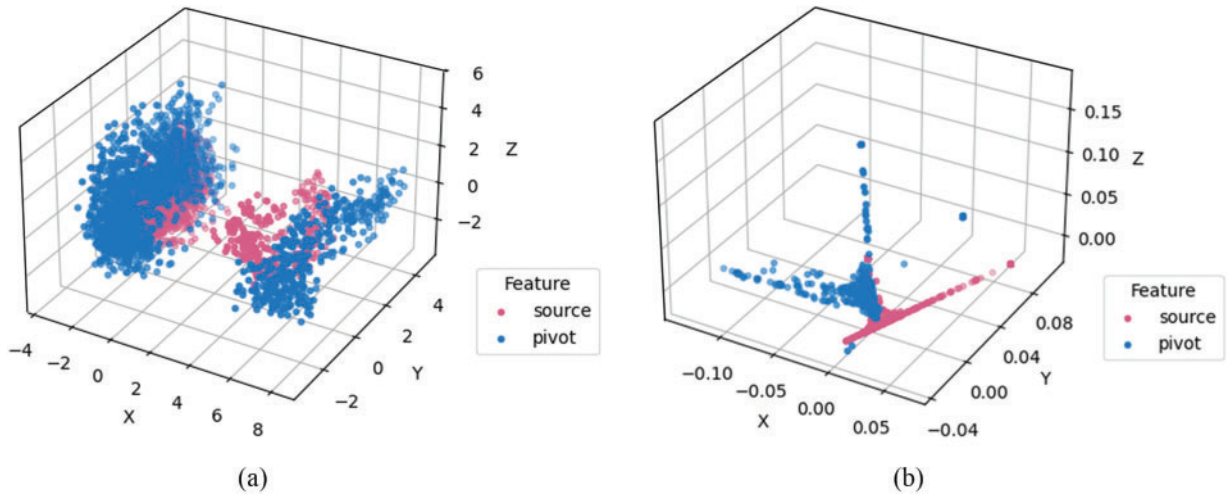

Figure 2: Feature distributions of source language (German) and pivot language (English) after dimensionality reduction with PCA and LLE. (**a**) Source and pivot feature distribution after PCA. (**b**) Source and pivot feature distribution after LLE

NMTs learn representations of different languages from large-scale corpus, and the vector representations of sentences with the same semantics across languages exhibit similarities. This similarity enables the translation of zero-resource language pairs by converting source language representations into pivot language representations. Kim et al. [12] proposed a linear pivot adapter to adapt the source language encoder output of Ms2p with the pivot language encoder output of Mp2t. The adapter was trained by minimizing the distance between these outputs, but it demonstrates limited performance in zero-shot scenarios. Zhang and Li [19] proposed a transfer-learning-based approach that encourages auxiliary models to learn representations within the same pivot language space during the pre-training, relying solely on a parameter freezing mechanism. In contrast to previous works, we pre-defined the representation space such that the pivot language spaces of Ms2t and Mp2t are similar. And then, we use a transformation method to achieve the spatial transformer. Specifically, our approach first utilizes PLM to establish a unified pivot representation space. During pre-training, we enforce adherence to this space by initializing and freezing the parameters of the Ms2p decoder and Mp2t encoder. Subsequently, we construct the final source-target translation model using the pre-trained encoder of Ms2p and decoder of Mp2t. To enhance feature migration, we integrate a feature converter between the final encoder and decoder and fine-tune the model using a synthetic parallel corpus.

3 Approach

Our focus is on machine translation of zero-resource language pairs, where no parallel corpus is available during training. The pivot-based approach, which involves a third language (such as English) provides a versatile solution. However, we propose a shared space transfer method that utilizes abundant pivot monolingual corpora to train a pivot PLM and uses its parameters to initialize auxiliary Ms2p and Mp2t models. To enhance the feature migration, we incorporate a feature converter between the final Ms2t model, consisting of the Ms2p encoder and the Mp2t decoder. The feature converter uses synthetic parallel corpus for end-to-end training. We will describe our approach in detail in this section.

3.1 Model Architecture

The simple pivot-based NMT approach underutilizes pivot monolingual corpora and then may introduce model bias due to Ms2p and Mp2t being trained by independent parallel corpora. Intuitively, we define the representation space of pivot language using PLM and train translation models (Ms2p and Mp2t) within this representation space. We illustrate our method in Fig. 3, which is divided into three stages. Let S, P, and T represent the source, pivot, and target languages, respectively.

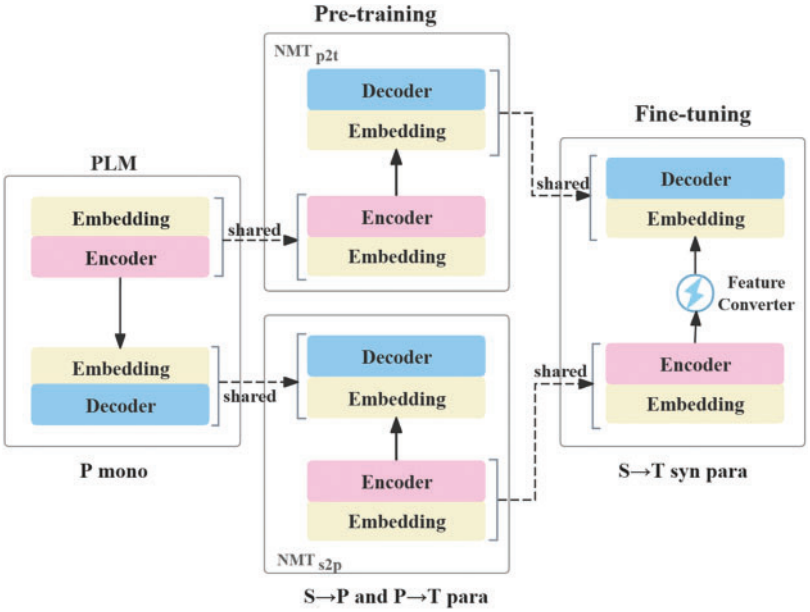

Figure 3: Shared space transfer method. The dashed line indicates the parameter's initialization. The present model shares the representation space during the pre-training and fine-tuning phase by initializing all parameters and freezing part parameters

We utilize corpora from S and P parallel corpora to train the PLM, which is then used to initialize the encoder parameters of Ms2p and the decoder parameters of Mp2t during the pre-trained stage. To ensure that Ms2p and Mp2t remain aligned with the representation space defined by P, we freeze specific parameters during training. We introduce a feature converter between the encoder outputs of Ms2p and Mp2t during fine-tuning. The final model of Ms2t, combining the Ms2p encoder, feature converter, and Mp2t decoder, is fine-tuned using limited synthetic parallel data. The feature converter adapts representations to ensure compatibility with the target decoder, even when models operate in similar representation spaces.

In experiments, we utilize BART [41] as a pivot language pre-trained language model (PLM), following a standard Transformer structure [42]. Both Ms2p and Ms2t use Transformers with $N = 6$ identical layers, in their encoder/decoder stacks, as shown in Fig. 4. Each encoder layer includes two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected network. After residual connections are performed between these sub-layers, layer normalization is followed. Similar to the encoder layer, but decoder layer with an extra sub-layer that executes multi-head attention on encoded representation.

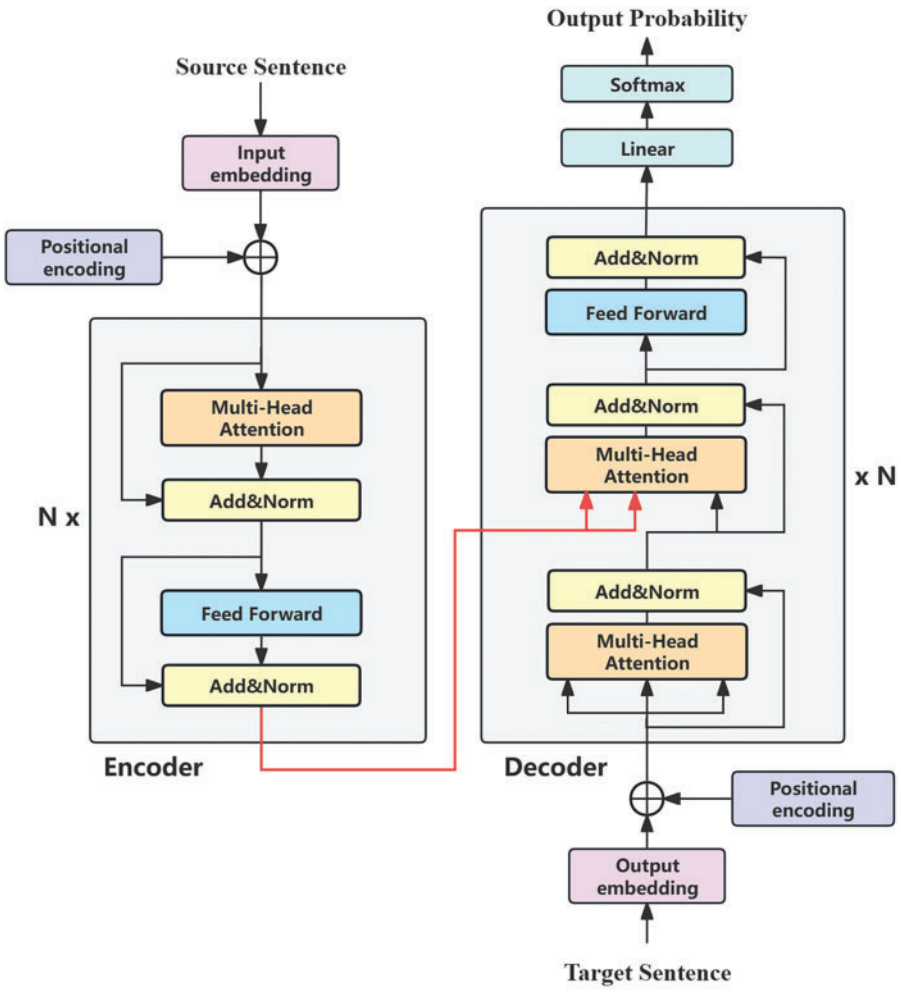

Figure 4: Transformer architecture

To predict the next token, the Transformer learns a linear transformation and a SoftMax function that converts the decoded feature representation to probabilities of token distribution. Additionally, to address the neglect of location information by attention, the Transformer proposes "position encoding" to encode position information using sine and cosine functions, which are then added to the input/output embedding.

Directly using the PLM's encoder/decoder for Ms2p and Mp2t without fine-tuning yields poor translation performance. Even for identical languages, representation spaces diverge when models are trained on different data. To mitigate this, we separately train Ms2p and Mp2t on their respective parallel corpora, further decreasing the discrepancy between the two models. Then, we freeze PLM-initialized parameters in lower layers (details below) to maintain effective translation in a similar representation space as much as possible. We finally share pivot vocabulary across models. Guided by findings that lower Transformer layers capture lexical features while higher layers encode semantics [43], and considering the sensitivity of cross-attention to pruning [44,45], we adopt the following freezing strategies:

Decoder Ms2p decoder initialized by PLM decoder. The embedding layer and self-attention of the lowest 3 layers are frozen, remaining parameters, including layer normalization and cross-attention, are fine-tuned.

Encoder Mp2t encoder initialized by PLM encoder. The embedding layer and self-attention mechanisms of the lowest 3 layers are frozen, and other parameters including layer normalization are fine-tuned.

This strategy encourages shared representation spaces while accommodating residual discrepancies. The feature converter bridges these gaps, enabling the target decoder to effectively decode adapted source representations.

3.2 Feature Converter

In NMT, feature vectors satisfying the distributional hypothesis to represent textual data is a crucial aspect, as they encode semantic information about words. The encoder maps source language words into these feature vectors, which the decoder then translates into the target language's representation space.

Assuming Ms2p and Mp2t are well-trained models, minimizing deviations between their encoded features is essential for effective decoding. When the encoding representation for sentences with the same semantics from the Ms2p encoder is similar to that from the Mp2t encode, the Mp2t decoder can decode them into target sentences. To achieve this, we propose a feature converter (see Fig. 5) that transforms the output of the Ms2p encoder into a representation compatible with the Mp2t decoder. Since both models share a pivot-language representation space defined by the pre-trained PLM, their spaces exhibit inherent similarities. Semantic equivalence between languages corresponds to analogous vector positions in these spaces, enabling direct conversion of source vectors into target representations.

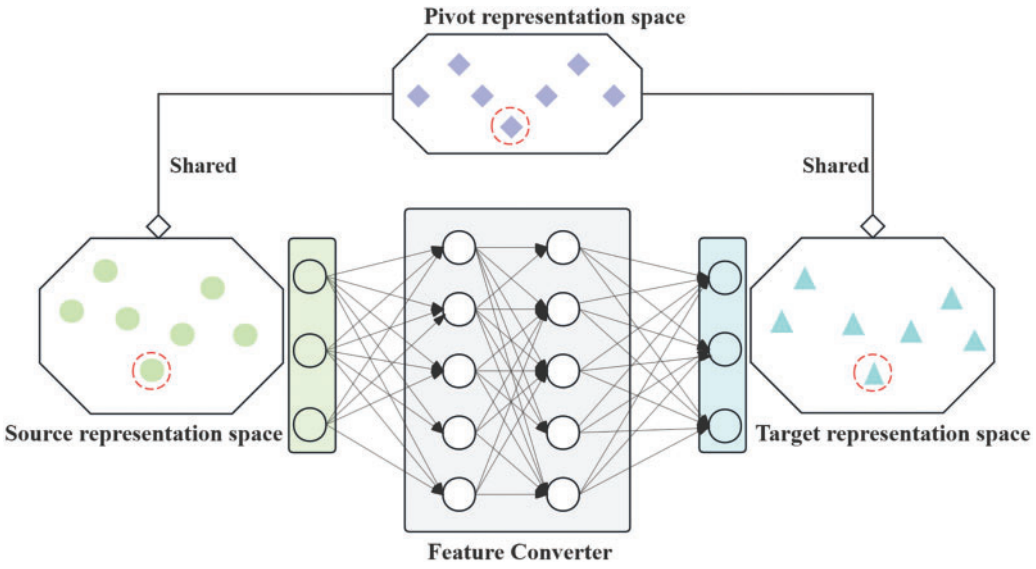

Figure 5: Feature converter with shared pivot space. Source language and target language learn their respective representation space under the constraint of shared pivot representation space

Shared representation space. During training, Ms2p and Mp2t are trained separately using the freezing strategy outlined in Section 3.1. Specifically, the encoder and decoder of pivot PLM initialize the encoder of Mp2t and the decoder of Ms2p. Then, the embedding and the lowest three layers of the Ms2p decoder and Mp2t encoder were frozen separately during training.

Synthesize parallel data. Synthetic data can be generated by directly translating the pivot sentence to the target sentence, or training a pivot→source model to back-translate the pivot sentence to the source

sentence. The former approach can be regarded as a teacher-student method, where the pivot→target model (teacher) guides the source→target model (student).

Training feature converter. We directly apply the feature converter between the source encoder and the target decoder. The feature converter transforms the source language representation space into pivot language representation space. We froze the encoder and decoder of the final source→target model and used synthetic parallel corpus to train feature converter via maximizing likelihood estimation.

The final source→target model combines the Ms2p encoder, feature converter, and Mp2t decoder. Due to linguistic divergences (e.g., syntax, vocabulary, word order), transformations between source and pivot spaces are not strictly one-to-one. The semantic correspondence between languages often presents non-linear geometric distortion. The deep network requires techniques (e.g., residual connection) to resolve gradient vanishing/exploding. This means complex non-linear transformations may introduce uncontrollable semantic drift (e.g., excessive transform destroys language structure). Thus, we employ a two-layer feed-forward network as the feature converter, with each layer consisting of a linear mapping followed by a ReLU activation function as:

$$F(x) = \max\left(\left(0, W_1 x + b_1\right) W_2 + b_2\right) \tag{3}$$

where W_1 and W_2 are linear mapping weights and b_1, b_2 are biases. The key to our method is to convert the source feature vector into the target representation space. This differs from the linear pivot adapter [12], which is trained independently by minimizing the distance from the source representation to the pivot representation. In our translation model, we directly end-to-end train the feature converter. In specific settings, we have found that a non-linear feature converter outperforms a linear feature converter. We have also observed that deeper feed-forward networks fail to achieve better performance in this context. Moreover, we use a consistent tokenization scheme that learns from a jointed corpus of the pivot language when pre-training Ms2p and Mp2t to mitigate the misalignment of sub-word tokenization resulting from different corpora.

4 Experiments

4.1 Dataset

We evaluated the present method on multiple language pairs, including synthetic zero-resource pairs, German to French (De→Fr), German to Czech (De→Cs), and Turkish to Hindi (Tr→Hi) with English (En) as the pivot language, and the real zero-resource language pair, Mongolian to Vietnamese (Mn→Vi) with Chinese (Ch) as the pivot. These language pairs lack direct source-target parallel data, they do have ample source-pivot and pivot-target parallel corpora. Data sources included WMT 2018, WMT 2019, and OPUS. For Mn→Vi, we constructed the test data via Google Translate and professional translators due to limited public corpora. To train the feature converter and comparative models, we synthesized the source-target parallel data using pivot→source or pivot→target models. The experiment data statistics are provided in Table 1, where the 'syn' label indicates that the parallel corpus is synthetic.

Table 1: Training data statistics in all experiments

	Train	Valid	Test
De→En	160,293	7283	6750
En→Fr	168,167	7643	4493
De→Fr	53,408 (syn)	2424 (syn)	2896

(Continued)

Table 1 (continued)

	Train	Valid	Test
De→En	182,957	8316	9106
En→Cs	90,214	4100	5716
De→Cs	42,800 (syn)	2048 (syn)	2359
Tr→En	483,452	27,016	26,985
En→Hi	481,881	26,545	26,705
Tr→Hi	60,235 (syn)	3318 (syn)	2102
Mn→Ch	328,799	18,312	18,263
Ch→Vi	126,332	6910	7021
Mn→Vi	48,723 (syn)	1984 (syn)	1822

Note: The 'syn' label indicates that the parallel corpus is synthetic.

Regarding the synthetic parallel data, we adopt the back-translation method from either the pivot→target or pivot→source model. Table 2 shows the performance of pre-training NMT models. For the De→Fr and Mn→Vi synthetic parallel data, we utilized En→Fr and Ch→Mn models, which can achieve good performance both on BLEU and COMET to generate French and Mongolian sentences corresponding to the De→En and Mn→Ch parallel corpus. For De→Cs and Tr→Hi pairs, we chose En→Cs and En→Tr models to synthesize the parallel corpora with a lower BLEU but higher COMET. Taking the synthesis of De-Fr parallel data as an example, we first randomly selected De-En parallel sentences from De→En training and valid dataset and employed En→Fr NMT to translate En sentences in these pairs into Fr sentences. Finally, we synthesized the De-Fr parallel corpus by aligning the De and Fr sentences corresponding to the same En sentences.

Table 2: The performance of pre-training NMT models

		BLEU	COMET
De→Fr (syn)	De→En	34.87	0.78
	En→De	33.57	0.73
	En→Fr	39.71	0.79
De→Cs (syn)	De→En	33.04	0.73
	En→De	28.54	0.69
	En→Cs	17.07	0.75
Tr→Hi (syn)	Tr→En	36.63	0.80
	En→Tr	33.44	0.87
	En→Hi	34.86	0.77
Mn→Vi (syn)	Mn→Ch	35.46	0.83
	Ch→Mn	37.36	0.86
	Ch→Vi	25.52	0.75

Note: The underlined NMT models that we have used to synthesize parallel corpus for zero-resource language pairs.

4.2 Model Training

We implemented all algorithms using the PyTorch framework and followed the pre-processing steps provided by the Fairseq toolkit [46]. Our model is based on the standard Transformer with 8 attention heads in each encoder and decoder sub-layer. We performed tokenization and true-casing on all corpora using Moses, and applied byte-pair encoding (BPE) [47] with 32,000 merge operations for pre-processing. The word embedding size is 512, and the maximum token count is set to 4096 tokens in each batch size. For optimization, Adam optimizes [48] with an initial learning rate $\beta = (0.9, 0.98)$ and label smoothing is 0.1. The dropout rate was 0.1 for each model. During inference, we used a beam size of 5 for greedy search. In experiments, we evaluated the translation performance using BLEU[1] [49], TER [50], and COMET [51] as primary metrics. COMET takes advantage of the pre-trained multilingual language model to evaluate the translation quality with a higher correlation than human judgment. We used the wmt22-comet-da[2] model to evaluate COMET scores.

To establish a shared representation space centered on the pivot language, we integrated pivot corpus from both source-pivot and pivot-target parallel corpora, which is subsequently utilized for PLM training. During the training of both NMTs2p and NMTp2t, we share the subword vocabulary of pivot language between PLM and NMTs.

4.3 Baselines

We address language pair translation in a zero-shot scenario using a pivot language. To validate the effectiveness, we compare it with the following baselines, each corresponding to one of three categories of pivot-based NMTs:

Direct pivot baseline: The baseline method involved training Ms2p and Mp2t models separately with parallel corpora, and then translating the source sentence to the target sentence by piping the output of Ms2p into the Mp2t model. This method represents a strong baseline.

Teacher-Student baseline (T-S): The baseline method proposed by Chen et al. [52] considers the pivot→source model as a "teacher" to generate the source→target data. The synthetic parallel corpus of the source-target is then used to train a "student" model, i.e., the source-target model. In our experiments, both the teacher and the student models adopt the Transformer model.

Step-wise Pre-training baseline (SP): The baseline method proposed by [12] involves training the target decoder to align with the source language's representation space through three stages. Firstly, train an Ms2p model using a joint vocabulary. Then, train an Mp2t model, initializing its encoder with the frozen Ms2p encoder. Finally, fine-tune the Ms2t model on the synthetic parallel corpus, initializing it with the Ms2p encoder and Mp2t decoder.

4.4 Result

Training analysis. Our method fine-tunes a feature converter while freezing the encoder and decoder initialized by auxiliary Ms2p and Mp2t. We present the train and valid loss for the teacher-student and our method in Fig. 6. Our method converges faster since only the converter's parameters are updated. Despite slower training loss reduction, the lower validation loss demonstrates stable training and robust performance.

[1] BLEU signature: BLEU+c.mixed+l.{de-fr, de-cs, tr-hi, mn-vi}+#.1+s.exp+tok.13a+v.2.4.0.
[2] https://huggingface.co/Unbabel/wmt22-comet-da (accessed on 27 May 2025).

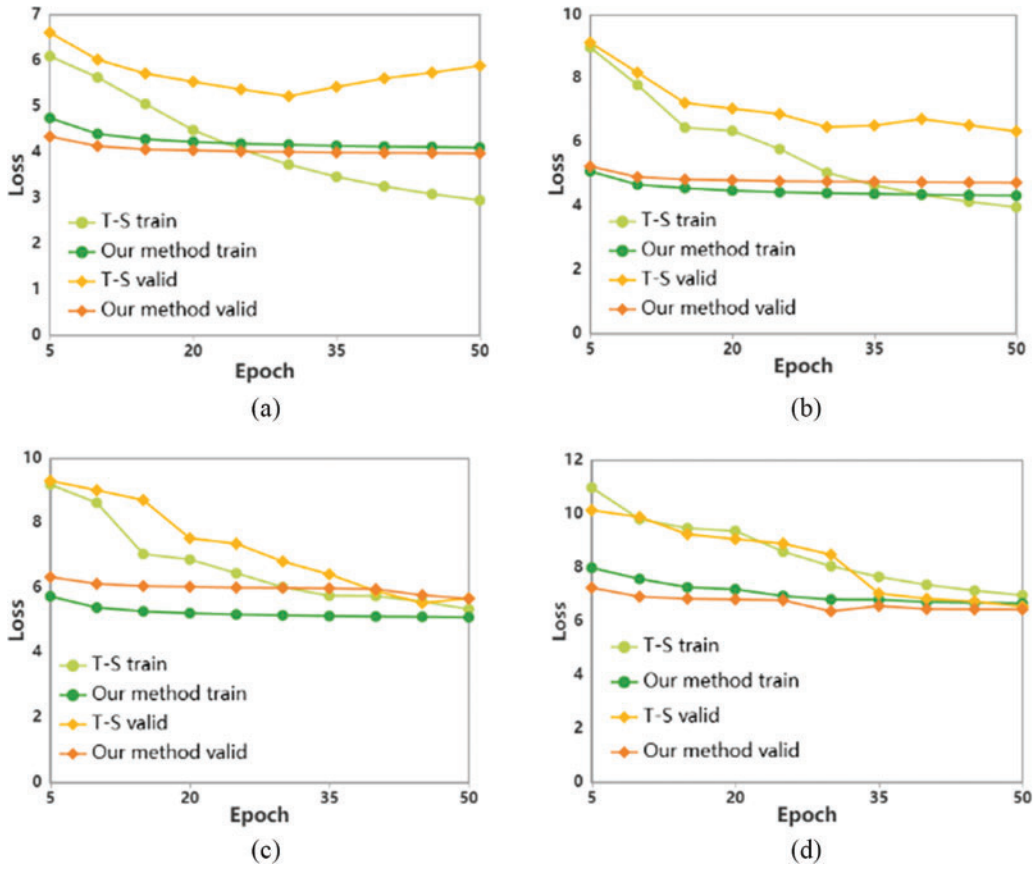

Figure 6: Train and valid loss curve in the training stage. (**a**) De→Fr loss curve; (**b**) De→Cs loss curve; (**c**) Tr→Hi loss curve; (**d**) Mn→Vi loss curve

In Tables 3–6, we present the performance of our model and baseline results for De→Fr, De→Cs, Tr→Hi and Mn→Vi. While the direct pivot method achieves strong performance, its dual-decoder architecture incurs significant computational overhead. Our method narrows this gap, trailing the direct pivot by no more than 2 BLEU while achieving the lowest TER and best COMET score for De→Fr. The step-wise pre-training baseline initially underperforms but improves substantially after fine-tuning with synthetic source-target parallel corpora, exceeding 10 BLEU across all language pairs. When combined with our feature converter, we can achieve the best COMET scores for De→Cs, Tr→Hi, and Mn→Vi. This demonstrates the feature converter's efficacy in preserving source semantics. The performance of the teacher-student model depends on the translation quality of the pivot-target model. For example, the noisy De→Cs synthetic parallel corpus (due to poor En→Cs translation) results in a 15.69 BLEU score drop compared to De→Fr. Initializing the teacher-student model with the Ms2p encoder and Mp2t decoder, both pre-trained with a pivot PLM, can enhance performance.

Table 3: The evaluation on the De→Fr language pairs with En as a pivot language

Method	BLEU	TER	COMET
Direct pivot	22.14	66.53	0.69
T-S	19.38	69.89	0.46

(Continued)

Table 3 (continued)

Method	BLEU	TER	COMET
T-S + PLM	20.69	68.12	0.48
SP	1.62	–	0.23
SP + Synthetic data	16.75	73.67	0.63
SP + Feature converter	17.48	70.26	0.64
Shared space transfer	**21.53**	**65.84**	**0.70**

Note: T-S is for the teacher-student model and SP is for the step-wise pre-training. "-" denotes the TER exceeds 100. The bold values denote the best records in simulations.

Table 4: The evaluation of the De→Cs language pairs with En as the pivot language.

Method	BLEU	TER	COMET
Direct pivot	15.46	76.57	0.68
T-S	3.69	–	0.37
T-S + PLM	4.42	–	0.35
SP	2.22	–	0.31
SP + Synthetic data	10.49	92.87	0.62
SP + Feature converter	11.56	91.32	**0.63**
Shared space transfer	**13.37**	**75.02**	0.58

Table 5: The evaluation of the Tr→Hi language pairs with En as the pivot language

Method	BLEU	TER	COMET
Direct pivot	21.10	61.05	0.57
T-S	15.01	85.79	0.42
T-S + PLM	16.28	87.26	0.44
SP	4.30	–	0.28
SP + Synthetic data	18.87	73.05	**0.56**
SP + Feature converter	**19.64**	76.31	0.55
Shared space transfer	19.05	**61.05**	0.51

Table 6: The evaluation on the Mn→Vi language pair with Ch as a pivot language

Method	BLEU	TER	COMET
Direct pivot	25.43	88.81	0.56
T-S	15.32	–	0.47
T-S + PLM	16.44	95.21	0.44
SP	5.90	–	0.41
SP + Synthetic data	22.90	74.61	0.57
SP + Feature converter	23.56	71.23	**0.59**
Shared space transfer	**23.98**	**77.93**	0.56

The direct pivot method, while effective, suffers from error propagation due to its dual-decoder architecture and challenges in cascade training. All existing approaches prioritize direct source→target translation models when both Ms2p (source→pivot) and Mp2t (pivot→target) exhibit strong performance. However, the teacher-student method underperforms if the teacher model (e.g., pivot→target) is low-quality. By defining the source language representation space via the pivot PLM, we ensure the compatibility between the source space and the target decoder. Notably, our feature converter can be integrated into other frameworks to enhance their performance. Experimental results confirm that our method outperforms the baseline models in BLEU and obtains the best COMET score for De→Fr. The step-wise pre-training with our feature converter can achieve better performance for De→Cs, Tr→Hi, and Mn→Vi language pairs.

Translation quality. To rigorously compare translation fidelity, we benchmark against SP+Feature Converter (step-wise pre-training baseline). As shown in Table 7, our translation achieves the highest COMET score, indicating its translation quality is closer to human-judged quality. Additionally, our method is within 1 BLEU score of direct translation in BLEU, while matching its TER. The step-wise pre-training model has a lower TER 28.58 than the teacher-student model, and the translation quality is inferior. The teacher-student model accurately translates "online-spielen" into "jouer en ligne", whereas the step-wise pre-training model translated it to "en ligne" and missed the word "derzeit".

Although our method did not achieve the highest BLEU score, it achieved the highest COMET score. We noticed that the translated sentences produced by the direct pivot and teacher-student methods tend to include more words, increasing the likelihood of mistranslation for the step-wise pre-training model. Our method strikes a good balance between the translation quality and similarity to the reference translation.

Freeze strategy influence: We analyze different pivot PLM freeze strategies in the De-En and En-Fr translation models. In our experiments, we initialize the Ms2p encoder and Mp2t decoder by using pivot PLM and train them according to the freeze strategy. We set freezing strategies for the decoder and encoder respectively according to the layer-wise and component-wise freezing. The details are as follows:

No-Frozen: Initialize the De→En decoder and En→Fr encoder by using pivot (English) PLM. All parameters can be fine-tuned in training time.

Embedding-Frozen: After initializing the encoder or decoder with pivot PLM, we freeze the embedding layer while other parameters are fine-tuned.

Lower-Frozen: After initializing the encoder or decoder with pivot PLM, the attention parameters of the lowest three layers and embedding layer are frozen. For the De→En decoder, the cross-attention of frozen layers is fine-tuned.

All-Frozen: After initializing the encoder or decoder with the pivot PLM, the attention parameters of all layers are frozen. For the De→En decoder, the cross-attention of frozen layers is fine-tuned.

Table 7: An example of pivot-based NMTs for the De→Fr language pair

Method	Source	derzeit verbringen wir 3 milliarden stunden pro woche mit online-spielen.
	Pivot	right now we spend three billion hours a week playing online games.
	Target	Aujourd'hui, nous passons trois milliards d'heures par semaine à jouer en ligne.
Direct pivot	Pivot	we currently spend three billion hours a week playing online games.
	Target	nous passons actuellement 3 milliards d'heures par semaine à jouer aux jeux en ligne. (BLEU: 44.80, TER: 42.86, COMET: 0.94)
SP + Feature converter	Target	nous passons 3 milliards d'heures par semaine en ligne. (BLEU: 28.38, TER: 35.71, COMET: 0.89)
T-S	Target	en ce moment, nous passons à peu près 3 milliards d'heures par semaine à faire des jouer en ligne. (BLEU: 34.11, TER: 64.29, COMET: 0.85)
Shared space transfer	Target	en ce moment, nous passons 3 milliards de heures par semaine à jouer en ligne. (BLEU:43.85, TER: 42.86, COMET: 0.93)

In Table 8, we compare the model that does not utilize pivot PLM initialization. For De-En and En-Fr, using PLM initialization without any freezing strategy (No-Frozen) achieves BLEU scores of 35.01 and 41.27, respectively, outperforming direct models. This indicates that the PLM's representations enhance translation performance. Our results show that different freezing strategies affect performance: the more layers that are frozen, the worse the performance tends to be. Specifically, using an all-frozen strategy for the decoder in the De→En translation results in a BLEU score drop over 20, while En→Fr only achieves 8.15 BLEU. In contrast, lower freezing strategies for the De-En and En-Fr maintain good performance, with differences from the best BLEU scores of no more than 2 and 5 points, respectively. Compared with the No-Frozen strategy, varying degrees of performance degradation are observed across different freezing strategies where the frozen layers discontinue parameter updates. Embedding-Frozen and Lower-Frozen strategies only decrease 0.14 BLEU and 0.85 BLEU on De→En, as both PLM and NMT are more inclined towards representing lexical and syntactic features at lower layers, meanwhile, higher layer learn task-specific representation. Thus, the All-Frozen strategy limits high layers to learn translation-specific patterns, which dramatically reduces performance, decreasing by 27.65 BLEU and 33.12 BLEU on De→En and En→Fr. Note that freezing only the embedding layer can enhance translation performance, but this approach may deviate from the representation space defined by the pivot language.

The lower layers of the language model mainly encode lexical and local syntactic features of the target language, while higher layers focus on semantic integration and task-specific representations [53]. Freezing these lower layers preserves general linguistic knowledge and avoids damaging the modeling ability of pre-trained language models [54] while allowing fine-tuning of higher layers for task adaptation. For translation tasks, the Transformer decoder needs to simultaneously learn target language generation and cross-language

alignment [55]. An all-frozen strategy prevents the Transformer from achieving cross-language alignment, severely degrading performance. Thus, while freezing all encoder or decoder layers preserves the pivot representation space effectively, it sacrifices translation quality. Our lower-freezing strategy balances this trade-off: retaining pivot-space alignment while enabling sufficient flexibility for task-specific adaptation.

Table 8: The results for different freezing strategies

Freezing strategy	BLEU (De→En)	BLEU (En→Fr)
Direct	33.57	40.14
No-Frozen	35.01	41.27
Embedding-Frozen	34.87	40.02
Lower-Frozen	34.16	34.87
All-Frozen	7.36	8.15

Converter efficiency. We analyzed the impact of feature converter design and different sizes of synthetic data on translation performance. Auxiliary Ms2p and Mp2t models are trained by using a Lower-Frozen freezing strategy with the pivot PLM. Despite shared initialization, Ms2p and Mp2t occupy distinct representation spaces. Directly combining their encoder and decoder ("plain transfer") yields poor performance (<2 BLEU). Our approach adds the extra feature converter between Ms2p encode and Mp2t decoder, and fine-tunes the model by synthetic parallel corpus. Our feature converter transforms the representation of the Ms2p encoder into the Mp2t decoder. Fig. 7 shows the performance of the final model with different feature converters and synthetic data. With the same synthetic parallel data, the non-linear converter performs better than the linear converter. For De→Fr, the non-linear converter is higher than linear converters 1.49, 1.67, and 1.21 BLEU. The fine-tuning data ranging from 10k, 50k to 160k can improve 2 BLEU and 0.13 for De→Fr with a non-linear converter.

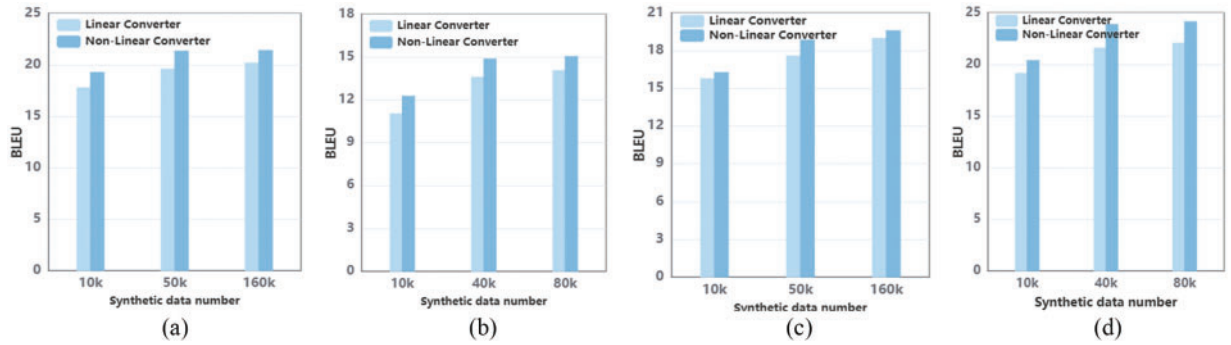

Figure 7: The performance of different synthetic data numbers and converters. (**a**) De→Fr BLEU; (**b**) De→Cs BLEU; (**c**) Tr→Hi BLEU; (**d**) Mn→Vi BLEU

We pre-trained the auxiliary Ms2p and Mp2t with the freezing strategy, but there is still a deviation between the representation space of the source and pivot language. The performance for plain transfer is poor. The feature converter can narrow the gap effectively, but there is a small amount of synthetic parallel data that can obtain positive performance. The results indicate the non-linear converter is suitable for feature transformation from source to pivot representation. However, even with a significant increase in fine-tuning

data, the feature converter's improvement is limited because of its simple network structure. This intrigues us to build a feature converter with more stronger transformation ability in the future.

Language model effect. We compared the proposed method using BART and RoBERTa [56] as LMs for De→Cs and Mn→Ch. The results are shown in Table 9. Using BART as a pivot outperforms RoBERTa in both pre-trained Ms2p and Mp2t and the final Ms2t model with feature converter. Compared to RoBERTa, the BLEU is higher at 2.1 and 1.24 respectively on De→Cs and Mn→Ch. BART adopts the encoder-decoder architecture, which maintains compatibility with standard Transformer, and its pre-training paradigm combines bidirectional and autoregressive objectives, which effectively handle text generation tasks. However, as RoBERTa has used an encoder-only framework, it limited text generation capabilities.

Table 9: Comparison of our method using BART and RoBERTa as LMs

		BART		RoBERTa	
		BLEU	**COMET**	**BLEU**	**COMET**
De→Cs	De→En	34.87	0.78	33.95	0.76
	En→Cs	17.07	0.75	17.48	0.79
	De→Cs	13.37	0.58	11.27	0.53
Mn→Vi	Mn→Ch	35.46	0.83	34.02	0.79
	Ch→Vi	25.52	0.75	24.36	0.75
	Mn→Vi	23.98	0.56	22.74	0.51

5 Conclusion

Our work introduced a novel pivot-based approach to enhance neural machine translation for zero-resource language pairs by aligning cross-lingual representation spaces. We analyzed the similarity of language representation across different languages and proposed a strategy involving parameter initialization and freezing to share the pivot language space between the Ms2p and Mp2t models. Furthermore, we proposed a feature converter to ensure the conversion between source language space and pivot language space. Experiment results on four language pairs including both synthetic and real zero-resource translation indicate the effectiveness of the present method. However, the COMET scores for these zero-resource pairs remain space for improvement as the present method cannot fully resolve word alignment issues. For future work, it is interesting to explore the integration of additional semantic information (e.g., part-of-speech tagging) and develop a more robust feature converter to enhance performance. Another method is to investigate leveraging LLMs with high-resource auxiliary languages to bridge the gap between the rigid literalness of NMT and the flexibility of human translation.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Natural Science Foundation of China (Grant number: Nos. 62172341 and 12204386), Sichuan Natural Science Foundation (Grant number: No. 2024NSFSC1375), Youth Foundation of Inner Mongolia Natural Science Foundation (Grant number: No. 2024QN06017), Basic Scientific Research Business Fee Project for Universities in Inner Mongolia (Grant number: No. 0406082215).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and design, Lingfang Li; data collection, Lingfang Li, Weijian Hu; draft manuscript preparation: Lingfang Li, Weijian Hu, Mingxing

Luo; funding acquisition and project administration, Mingxing Luo. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in WMT2018, WMT2019 and OPUS at https://www.statmt.org/wmt19, https://www.statmt.org/wmt18 and https://opus.nlpl.eu/corpora (accessed on 19 January 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Charoenpornsawat P, Sornlertlamvanich V, Charoenporn T. Improving translation quality of rule-based machine translation. In: Proceedings of the COLING-02: Machine Translation in Asia; 2002 Sep 1; Taipei, Taiwan.
2. Lopez A. Statistical machine translation. ACM Comput Surv. 2008;40(3):1–49. doi:10.1145/1380584.1380586.
3. Dumebi OM. Machine translation approaches: issues and challenges. Int J Comput Sci Issues. 2014;11(5):159.
4. Wang H, Wu H, He Z, Huang L, Church KW. Progress in machine translation. Engineering. 2022;18(2):143–53. doi:10.1016/j.eng.2021.03.023.
5. Junczys-Dowmunt M, Dwojak T, Hoang H. Is neural machine translation ready for deployment? A case study on translation directions. In: Proceedings of the 13th International Conference on Spoken Language Translation; 2016 Dec 8–9; Seattle, WA, USA.
6. Garg A, Agarwal M. Machine translation: a literature review. arXiv:1901.01122. 2018.
7. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridge the gap between human and machine translation. arXiv:1609.08144. 2016.
8. Dabre R, Chu C, Kunchukuttan A. A survey of multilingual neural machine translation. ACM Comput Surv. 2020;53(5):1–38. doi:10.1145/3406095.
9. Koehn P, Knowles R. Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation; 2017 Aug 4; Vancouver, BC, Canada. doi:10.18653/v1/W17-3204.
10. Sennrich R, Zhang B. Revisiting low-resource neural machine translation: a case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. doi:10.18653/v1/P19-1021.
11. Ranathunga S, Lee E-SA, Skenduli MP, Shekhar R, Alam M, Kaur R. Neural machine translation for low-resource languages: a survey. ACM Comput Surv. 2023;55(11):1–37. doi:10.1145/3567592.
12. Kim Y, Petrov P, Khadivi S, Khadivi S, Ney H. Pivot-based transfer learning for neural machine translation between Non-English languages. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China. doi:10.18653/v1/D19-1080.
13. Dabre R, Imankulova A, Kaneko M. Simultaneous mult-pivot neural machine translation. arXiv:2104.07410. 2021.
14. Cheng Y, Liu Y, Yang Q, Sun M, Xu W. Neural machine translation with pivot languages. arXiv:1611.04928. 2017.
15. Tokarchuk E, Rosendahl J, Wang W, Petrushkov P, Lancewicki T, Khadivi S, et al. Towards reinforcement learning for pivot-based neural machine translation with non-autoregressive transformer. arXiv:2109.13097. 2021.
16. Lakew SM, Lotito QF, Negri M, Turchi M, Federico M. Improving zero-shot translation of low-resource languages. In: Proceedings of the 14th International Conference on Spoken Language Translation; 2017 Dec 14; Tokyo, Japan.
17. Heafield ACK. Zero-resource neural machine translation with monolingual pivot data. In: Proceedings of the 3rd Workshop on Neural Generation and Translation; 2019 Nov 4; Hong Kong, China. doi:10.18653/v1/D19-5610.
18. Mhaskar S, Bhattacharyya P. Pivot based transfer learning for neural machine translation: CFILT IITB @ WMT 2021 triangular MT. In: Proceedings of the Sixth Conference on Machine Translation; 2021 Nov 10–11; Online.
19. Zhang M, Li L. Triangular transfer: freezing the pivot for triangular machine translation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. doi:10.18653/v1/2022.acl-short.72.

20. Zoph B, Yuret D, May J, Knight K. Transfer learning for low-resource neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1–5; Austin, TX, USA. doi:10.18653/v1/D16-1163.
21. Zhang H, Chen K, Bai X, Li X, Xiang Y, Zhang M. Exploring translation mechanism of large language models. arXiv:2502.11806. 2025.
22. Li J, Zhou H, Huang S, Cheng S, Chen J. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. Trans Assoc Comput Linguist. 2014;12:576–92. doi:10.1162/tacl_a_00655.
23. Sizov F, España-Bonet C, Van Genabith J, Xie R. Analysing translation artifacts: a comparative study of LLMs, NMTs, and human translations. In: Proceedings of the Ninth Conference on Machine Translation; 2024 Nov 15–16; Miami, FL, USA. doi:10.18653/v1/2024.wmt-1.116.
24. Muennighoff N, Wang T, Sutawika L, Penedo G, Le Scao T, Melas-Kyriazi L, et al. Crosslingual generalization through multitask finetuning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, ON, Canada. doi:10.18653/v1/2023.acl-long.891.
25. Wendler C, Veselovsky V, Monea G, West R. Do llamas work in English? On the latent language of multilingual transformers. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand.
26. Zhu S, Cui M, Xiong D. Towards robust in-context learning for machine translation with large language models. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); 2024 May 20–25; Torino, Italia.
27. Arora A, Jurafsky D, Potts C, Goodman ND. Bayesian scaling laws for in-context learning. arXiv:2410.16531. 2024.
28. Zhu S, Pan L, Li B, Xiong D. LANDeRMT: dectecting and routing language-aware neurons for selectively finetuning LLMs to machine translation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 13–15; Bangkok, Thailand. doi:10.18653/v1/2024.acl-long.656.
29. Pan S, Tian Z, Ding L, Zheng H, Huang Z, Wen Z, et al. POMP: probability-driven meta-graph prompter for LLMs in low-resource unsupervised neural machine translation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 13–15; Bangkok, Thailand. doi:10.18653/v1/2024.acl-long.537.
30. Chen G, Ma S, Chen Y, Zhang D, Pan J, Wang W, et al. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. doi:10.18653/v1/2022.acl-long.12.
31. Eronen J, Ptaszynski M, Nowakowski K, Chia ZL. Improving polish to english neural machine translation with transfer learning: effects of data volume and language similarity. In: Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation; 2023 Jun 15; Tampere, Finland.
32. Fu Y, Hospedales TM, Xiang T. Transductive multi-view zero-shot learning. IEEE Trans Pattern Anal Mach Intell. 2015;37(11):2332–45. doi:10.1109/TPAMI.2015.2408354.
33. Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015 Sep 17–21; Lisbon, Portugal. doi:10.18653/v1/D15-1166.
34. Tan Z, Wang S, Yang Z, Chen G, Huang X, Sun M, et al. Neural machine translation: a review of methods, resources, and tools. AI Open. 2020;1:5–21. doi:10.1016/JAIOPEN.2020.11.001.
35. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany. doi:10.18653/v1/P16-1009.
36. Xia M, Kong X, Màrquez L. Generalized data augmentation for low-resource translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. doi:10.18653/v1/P19-1579.
37. Ko W, El-Kishky A, Renduchintala A, Chaudhary V, Goyal N, Guzmán F. Adapting high-resource nmt models to translate low-resource related languages without parallel data. In: Proceedings of the 59th Annual Meeting of

the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Online. doi:10.18653/v1/2021.acl-long.66.
38. Ji B, Zhang Z, Duan X, Zhang M, Chen B, Luo W. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York, NY, USA. doi:10.1609/aaai.v34i01.5341.
39. Cho K, van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of the SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation; 2014 Oct 25; Doha, Qatar. doi:10.3115/v1/W14-4012.
40. Aitken K, Ramasesh VV, Cao Y, Maheswaranathan N. Understanding how encoder-decoder architectures attend. In: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Dec 6–14; Online.
41. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. doi:10.18653/v1/2020.acl-main.703.
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17); 2017 Dec 4–9; Long Beach, CA, USA.
43. Tiedemann ARJ. An analysis of encoder representations in transformer-based machine translation. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; 2018 Nov 1; Brussels, Belgium. doi:10.18653/v1/W18-5431.
44. Li X, Wang C, Tang Y, Tran C, Tang Y, Pino J, et al. Multilingual speech translation from efficient finetuning of pretrained models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Online. doi:10.18653/v1/2021.acl-long.68.
45. Gheini M, Ren X, May J. Cross-attention is all you need: adapting pretrained transformers for machine translation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 7–11; Online. doi:10.18653/v1/2021.emnlp-main.132.
46. Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, et al. FAIRSEQ: a fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics; 2019 Jun 2–7; Minneapolis, MN, USA. doi:10.18653/v1/N19-4009.
47. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany. doi:10.18653/v1/P16-1162.
48. Kingma DP, Ba JL. Adam: a method for stochastic optimization; 2014. arXiv:1412.6980. 2014.
49. Papineni K, Roukos S, Ward T. A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul 6–12; Philadelphia, PA, USA. doi:10.3115/1073083.1073135.
50. Snover M, Dorr B, Schwartz R, Micciulla L. A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers; 2006 Aug 8–12; Cambridge, MA, USA.
51. Rei R, Stewart C, Farinha AC, Lavie A. COMET: a neural framework for MT evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. doi:10.18653/v1/2020.emnlp-main.213.
52. Chen Y, Liu Y, Cheng Y. A teacher-student framework for zero-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017 Jul 30–Aug 4; Vancouver, BC, Canada. doi:10.18653/v1/P17-1176.
53. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA. doi:10.18653/v1/N19-1423.

54. Peters ME, Ruder S, Smith NA. To tune or not to tune? Adapting pretrained representations to diverse tasks. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019); 2019 Aug 2; Florence, Italy. doi:10.18653/v1/W19-4302.
55. Voita E, Sennrich R. Analyzing the source and target contributions to predictions in neural machine translation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Online. doi:10.18653/v1/2021.acl-long.91.
56. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.