Alexandria Engineering Journal 116 (2025) 633–643

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Original article

Enhancing distant low-resource neural machine translation with semantic pivot

Enchang Zhu [a,b], Yuxin Huang [b], Yantuan Xian [b], Junguo Zhu [a], Minghu Gao [a], Zhiqiang Yu [a,*]

[a] School of Mathematics and Computer Science, Yunnan Minzu University, Kunming, 650500, China
[b] School of Information Engineering and Automation, Kunming University of Science and Technology, 650500, China

ARTICLE INFO

Keywords:
Machine translation
Chinese-Lao
Pivot
Adapter
Similar linguistic feature

ABSTRACT

Prior work has proved that pivot-based method can boost the performance of neural machine translation (NMT). However, in low-resource scenarios, the efficient of pivot-based method is impaired severely due to data sparsity problem. As a typical low-resource language pair, Chinese-Lao NMT suffers the same performance dilemma. In addition, due to the significant linguistic gap between Chinese and Lao, some traditional and effective low-resource translation methods, such as introducing similarity external knowledge, sharing word space, and literal translation, are not suitable for the translation of this language pair. Fortunately, it is highly adaptable to pivot strategy, as there is a pivot language, Thai, which is highly similar to the target language Lao. Here, we propose a novel approach for incorporating similar linguistic features between Thai and Lao into the Chinese-Lao translation model. Firstly, an in-depth linguistic similarity analysis of Thai and Lao is conducted. Secondly, an elaborate pivot-based translation framework with KL adapter is applied. Experiments on the Chinese-Lao translation task show that our approach can help transfer more linguistic knowledges from the Chinese encoder to the Lao decoder via similar linguistic features, achieving substantial improvements compared to the baseline models.

1. Introduction

Chinese-Lao is a representative distant language pair of low-resource languages, and research on Chinese-Lao NMT in past years has been limited. Owing to the lack of parallel training data, researchers devoted to Chinese-Lao NMT have to focus on corpus preprocessing and monolingual language models [1,2]. However, with the increasing economic activities of the two countries, the demand for translation between Chinese and Lao has become imminent. Therefore, it is important to explore how to design an efficient Chinese-Lao translation method in which an efficient translation model can be trained on a small-scale corpus.

Chinese and Lao belong to different language families. Specifically, Chinese is a Sino-Tibetan language, whereas Lao belongs to the Tai-Kadai language family. The Chinese-Lao language pair is mutually unintelligible because of prodigious cross-lingual differences. Therefore, it is logical to choose an appropriate pivot language to overcome this dilemma. Empirically, the optimal pivot language for Chinese-Lao translation contains the following features: (1) data-scale adaption: the scale of the Chinese-pivot corpus and Chinese-Lao corpus could be imbalanced. Specifically, the former is generally larger than the latter, the pivot language should have the ability to handle the data imbalance; and (2) highly similar to the target language: the pivot language should have remarkable cross-lingual similarities with Lao. To be best, the language family of pivot language and Lao should be the same. To meet the above requirements, we empirically choose Thai as the pivot language. To the best of our knowledge, researchers have conducted exploratory works [3–6] on the utilization of language features, and the effectiveness of incorporating language features into low-resource distant NMT has also been preliminarily revealed. However, the de facto NMT methods often overlook in-depth analysis of linguistic feature. As is known to all, human language is a regularized description of semantics, therefore in-depth researches from a linguistic perspective can undoubtedly reveal the essence of language. Apart from this, in the past, the utilization of language features often focused on preliminarily discussions, with insufficient exploration of how to integrate language features into standard NMT frameworks. To this end, we aim to leverage similar features of the Thai-Lao language pair to improve the performance of the Chinese-Lao translation. The contributions of this paper are as follows:

* Corresponding author.
E-mail addresses: zhuenchang.kmust@foxmail.com (E. Zhu), huangyuxin2004@163.com (Y. Huang), xianyt@kust.edu.cn (Y. Xian), jg.zhu@kust.edu.cn (J. Zhu), gmhfc@ymu.edu.cn (M. Gao), yzqyt@ymu.edu.cn (Z. Yu).

https://doi.org/10.1016/j.aej.2024.12.073
Received 29 July 2024; Received in revised form 14 December 2024; Accepted 20 December 2024
Available online 9 January 2025

1. Linguistic analysis. First, we study the linguistic similarity of the Thai-Lao language pair. We then discuss the practicability of choosing Thai as the pivot language to leverage the linguistic similarity to improve the performance of the Chinese-Lao NMT.

2. Data augmentation. We design a simple but effective data augmentation strategy and apply it sequentially to the each stage of pivot translation to overcome the data imbalance existed in the Chinese-Lao NMT, this strategy is also applicable to other distant low-resource scenarios with pivot-target similarity.

3. Pivot-based NMT model. To utilize the augmented corpus and available small-scale Chinese-Lao parallel corpus, we propose a pivot-based translation model for the Chinese-Lao NMT. The core idea is to build a new KL-adapter based Chinese-Lao NMT model by regroup the Chinese encoder and Lao decoder exacted from pre-trained Chinese-Thai and Thai-Lao translation models, and then fine-tuning the newly generated model using a small-scale Chinese-Lao parallel corpus.

The remainder of this paper is organized as follows. Section 2 introduces the relevant works. Section 3 discusses the linguistic similarities of the Thai-Lao language pair. In Section 4, we elaborate on the proposed approach in terms of data augmentation and translation-model building. Section 5 reports the experimental setup and main results. Section 6 presents the conclusions of this work.

2. Related work

2.1. Pivot-based NMT

A series of pivot-based methods have been proposed to address the data sparsity dilemma in low-resource scenarios, the naivest of which is reusing the pre-trained source→pivot and pivot→target models directly, e.g., joint training method [7] for pivot-based NMT. The central idea of the work is to build a connection between the source pivot and pivot-target NMT models, allowing them to interact with each other during the training procedure. Apart from this, dynamical route is also a good idea for pivot-based NMT. Leng et al. [8] used an unsupervised pivot-based approach for distant-language translation, in which the source language can be translated to a distant language through a multiple-hop path. Specifically, a learning method (LTR) is proposed to drive the model to choose a translation route dynamically between the source and target. Kim et al. [9] leveraged three pre-training strategies to enhance the connection between the source, pivot, and target languages. Specifically, this work first obtains end-to-end translation models of different language pairs using iteration-wise training. Then, the study adopts a smooth adapter to connect the trained encoder, decoder, and cross-lingual encoder trained by autoencoding the pivot language. Li et al. [10] introduced a novel reference language-based model for unsupervised NMT. The basic idea is that the reference language shares only parallel sentences with the source language, but the parallel sentences still provide a pivot signal to reconstruct the training of the unsupervised NMT model through a reference agreement strategy. For the simultaneous translation scenario, Dabre et al. [11] proposed a multi-pivot NMT method, which first translates a source language to multiple pivots and then simultaneously translates the pivots into the target language. These works show significant improvements when the pivot strategy is applied to leverage trained rich-resource models to cope with the inefficiency problem in low-resource scenarios. Chen et al. [12] proposed SixT model, which leverages the multilingual pretrained encoder with a two-stage training schedule and gets further improvement with a position disentangled encoder and a capacity-enhanced decoder, and achieves better performance on multi translation benchmark compared with strong multilingual NMT baselines. To improve the connection during model training, Tokarchuk, E. et al. [13] proposed to train a pivot-based NMT system with the reinforcement learning (RL) approach, which utilizes a non-autoregressive transformer and present an end-to-end pivot-based integrated model, facilitating training on source-target data. In addition to the significant works aforementioned, semantic pivot approaches [14,15], i.e., enhances the semantics between the source and target languages through bridging, have also become a feasible approach to breaking the bottleneck of distant language NMT. However, approaches that rely on a pivot language are often plagued by the issue of error propagation, where errors in the initial source-to-pivot translation continue to influence the subsequent pivot-to-target translation. Part of the reason for this can be found in the inconsistency between source-pivot and pivot-target parallel corpora, which are typically only loosely connected or entirely unrelated. Moreover, the independent training of source-to-pivot and pivot-to-target translation models exacerbates the issue, enlarging the linguistic gap between the source and target languages. Therefore, how to bridge semantic barriers while using pivot methods is what we need to tackle in this paper.

2.2. Transfer-based NMT

Transfer learning was first introduced to NMT by Zoph et al. [16], which first trains a high-resource language pair model, then transfers the pre-trained parameters to the low-resource language pair model to initialize and constrain training. Zhang et al. [17] proposed a transfer learning based approach that utilizes all types of auxiliary data, which trains auxiliary source-pivot and pivot-target translation models. The approach initializes some parameters of the pivot side with a pre-trained language model and freezes them to encourage both translation models to work in the same pivot language space, so that they can be smoothly transferred to the source-target translation model. In terms of using vocabulary for transfer, Nguyen et al. [18] and Kocmi et al. [19] used shared subword vocabularies to work with more languages and help target language switches. To consider the dynamic requirements of the vocabulary, Lakew et al. [20] transferred knowledge from the source-pivot model to the pivot-target model by means of a shared dynamic vocabulary. Moreover, Kim et al. [21] proposed additional techniques to enable NMT transfer even without shared vocabularies. Jiang et al. [22] proposed lexical constraint mechanism to transfer-based NMT procedure. Apart from this, some representative works designed transfer algorithms based on mask substructure [23], k-Nearest-Neighbour [24], lexicon embedding [25] and likelihood [26], achieving significant performance improvements. Nevertheless, these works mainly adopt strategies that augment the training data or assist in transferring the parameters from the rich-resource models to the low-resource models, neglecting the way to utilize similar features between the pivot and target languages.

Some prominent works aforementioned have demonstrated the effectiveness of the comprehensive use of pivot and transfer learning. For pivot-based transfer model training, the limitation of past methods is that the source encoder is trained to be used for source-pivot translation task, while the target decoder is trained to adapt to pivot-target translation task—not of a source-target translation. To tackle this, traditional methods utilize parallel corpora for source target fine-tuning. However, in low-resource scenarios, the effectiveness of this fine-tuning is limited and cannot effectively address potential biases and inaccuracies. To address this problem, our method synthetically uses pivoted data augmentation based on linguistic similarity and pivot-based transfer model training to mitigate the inconsistency of the pre-training stages, Weakening the deviations and inaccuracies in the translation process.

3. Thai-Lao linguistic similarity

Thai-Lao is a tonal-based language pair belonging to the Tai-Kadai language family. Intuitively, these two languages have remarkable similarities in both speech and writing. In fact, for speech, Thai and Laotian people can communicate orally. For writing, the Thai-Lao language pair shares a considerable number of correlative words etymologically, and the sentences of the Thai-Lao language pair usually have basically consistent head-initial syntactic structures [27].

Fig. 1. Thai-Lao linguistic similarity.

Table 1
Linguistic similarity measured by position embedding.

Classification	Language pair	Similarity
Distant language	Japanese-English	53.4%
	Chinese-English	62.2%
Similar language	French-English	76.3%
	German-English	78.1%
	Thai-Lao	81.7%

3.1. Token shape

Thai and Lao words consist of abugida tokens. Moreover, sentence pairs composed of similar words are typically linguistically similar [28]. As shown in Fig. 1, the shapes of the words in the parallel sentences of Thai and Lao are highly similar (the difference is often only in the degree of curvature of the characters). Overall, the writing difference between Thai and Lao usually lies in the smoothness of the tokens and the writing of a few tokens.

3.2. Syntactic structure

In addition to token-level similarity analysis, we also investigate similarity from the perspective of syntactic structure. To obtain word alignment, we use the GIZA++ tool [29] to preprocess the sentences in the Thai-Lao ALT [30][1] subset. Based on the alignment results, we follow Isozaki et al. [29] to use Kendall's τ as a metric for the calculation of cross-lingual similarity between the two languages. Unlike Levenshtein distance, which adjusts two sentences to be identical, Kendall's τ focuses on the cost of word reordering, and in Isozaki et al. [31], it is measured by increasing pairs:

$$\tau = \frac{\text{count(increase\_pairs)}}{\text{count(all\_pairs)}} \times 2 - 1 \tag{1}$$

This specific value is related to syntactic structural similarity and word alignment. Fig. 2 shows Kendall's τ of the Thai-Lao language pair on the ALT dataset. Note that since the Kendall's τ is asymmetric, the distribution of Thai→Lao (Subgraph (a)) and Lao→Thai (Subgraph (b)) are different. It can be observed that the average value of Kendall's τ for the Thai-Lao language pair is about 0.73, which indicates that there are considerable similarities in the syntactic structure level. As we have observed, Thai and Lao have remarkable linguistic similarities in terms of token shape and syntactic structure. We assume that these similarities can help transfer information from the source language (Chinese) to Lao. Intuitively, Thai is a promising the pivot language for Chinese-Lao NMT, as far as we know, there is no existing pivot-based work for Chinese-Lao translation by using cross-lingual similarity.

3.3. Distributed representation

The NMT model uses a distributed representation to represent the input sentences. To investigate whether linguistic similarity exists in the distributed representation. We use the position embedding metric [32] to evaluate the linguistic similarity in the Thai-Lao language pair. As shown in Table 1, French-English and German-English are much more similar than distant languages (Japanese-English and Chinese-English language pairs). The main reason is that, the syntactic structure and grammatical rules vary significantly between Indo-European languages and Sino-Tibetan languages. The similarity (difference) persists after the distributed representation. Moreover, Table 1 shows that Thai-Lao has a higher similarity than French-English and German-English language pairs; therefore, it is practicable to choose Thai as the pivot language for Chinese-Lao translation.

[1] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/.

4. Translation model

In this section, we describe our proposed method. As shown in Fig. 3, the final desired Chinese-Lao translation model consists of a fine-tuned Chinese encoder and Lao decoder. To this end, we first train a couple of intermediary translation models, Chinese-Thai model and Thai-Lao model, subsequently, we extract the Chinese encoder and the Lao decoder from the pre-trained models, and combine the encoder and decoder with the help of a KL adapter as the final Chinese-Lao translation model. We describe the training of intermediary translation models in Section 4.1. The procedure for building the final Chinese-Lao translation model is described in Section 4.2.

4.1. Intermediary translation models

As described above, the scales of the Chinese-Thai (zh-th) and Chinese-Lao (zh-lo) corpora are imbalanced. The Chinese-Thai translation model can be trained on a relatively sufficient parallel corpus, including a certain-scale homemade corpus and small-scale corpus of an open ALT dataset. Given source Chinese and corresponding target Thai parallel data as inputs. The Chinese-Thai translation model can be represented as $P\left(th|zh;\theta_{zh\to th}\right)$, where $\theta_{zh\to th}$ is the model parameter set trained on parallel sentences $D_{zh,th}$ by the maximum likelihood estimation [33]:

$$\hat{\theta}_{zh\to th} = \underset{\theta_{zh\to th}}{\operatorname{argmax}} \left\{\ell\left(\theta_{zh\to th}\right)\right\} \tag{2}$$

where $\ell\left(\theta_{zh\to th}\right)$ denotes log-likelihood and can be formally described as:

$$\ell\left(\theta_{zh\to th}\right) = \sum_{zh,th\in D_{zh,th}} \log P\left(th \mid zh;\theta_{zh\to th}\right) \tag{3}$$

For the Thai-Lao translation model, we use the linguistic similarities of the Thai-Lao language pair to build a combined corpus for training. Specifically, we adopt the following simple strategy:

One part of the training corpus is the existing small-scale parallel dataset, which is denoted as $D_{th,lo} = \left\{\left\langle th^{(n)}, lo^{(n)}\right\rangle\right\}_{n=1}^{N}$, on which the maximum likelihood estimation (MLE) loss function is used to update the translation model by maximizing the log likelihood of translation. The MLE loss function can be described as:

$$\ell\left(\theta_{th\to lo}\right) = \sum_{th,lo\in D_{th,lo}} \log P\left(lo \mid th;\theta_{th\to lo}\right) \tag{4}$$

The other part of the corpus for training is pseudo-parallel generated by literal (word-to-word) translation. Literal translation directly translates Thai sentences D'_{th} in the Chinese-Thai homemade corpus $D'_{zh,th} = \{\langle zh^{(\mathrm{m})}, th^{(\mathrm{m})}\rangle\}_{m=1}^{M}$ into Lao sentences D'_{lo} on a Thai-Lao lexicon. Then, the translated Lao sentences and Thai source sentences are combined into the Thai-Lao pseudo parallel corpus $D'_{th,lo} = \{\langle th^{(m)}, lo^{(m)}\rangle\}_{m=1}^{M}$. The pseudo-parallel corpus contains only slight noise owing to the high similarity between the two languages. The training loss on the pseudo-parallel corpus can be formulated as:

$$\ell\left(\theta'_{th\to lo}\right) = \sum_{th,lo\in D'_{th,lo}} \log P\left(lo \mid th;\theta'_{th\to lo}\right) \tag{5}$$

However, in accordance with the real distribution of parallel data, we use a parameter α to balance the information proportion during training:

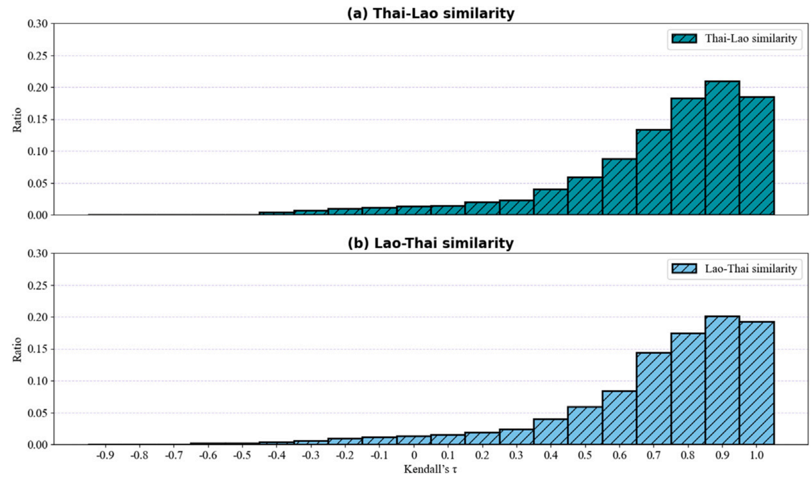

Fig. 2. Kendall's τ distribution of Thai-Lao language pair. Subfigure (a) represents Thai-Lao similarity, while Subfigure (b) denotes Lao-Thai similarity.

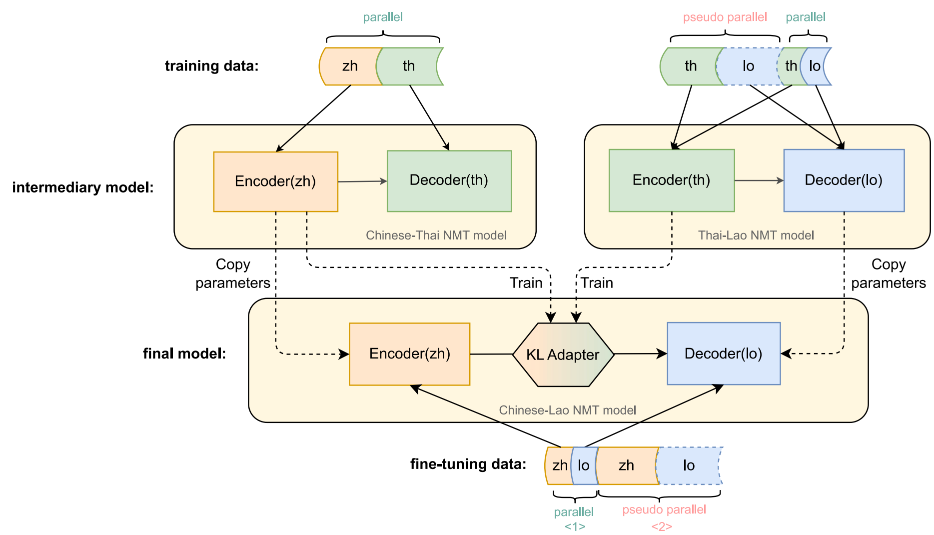

Fig. 3. Architecture of the pivot-based Chinese-Lao translation model. The intermediary models are the pre-trained source-pivot (Chinese-Thai) model and pivot-target (Thai-Lao) model, while the encoder and decoder of the final model are initialized separately from the pre-trained models through parameter copying. The training objective of KL adapter is to minimize the KL divergence between the probability distribution of the sentence representations generated by Chinese encoder and Thai encoder.

$$\ell\left(\theta''_{th\to lo}\right) = \alpha\ell\left(\theta_{th\to lo}\right) + (1-\alpha)\ell\left(\theta'_{th\to lo}\right) \tag{6}$$

4.2. Final Chinese-Lao translation model

Inspired by Kim et al. [9], we compose a new translation model using the Chinese encoder and the Lao decoder extracted from the Chinese-Thai NMT model and the Thai-Lao NMT model, respectively. The process can be formulated as follows:

$$encoder_{zh} = extractEnc\left(P\left(th \mid zh; \theta_{zh\to th}\right)\right) \tag{7}$$

$$decoder_{lo} = extractDec\left(P\left(lo \mid th; \theta_{th\to lo}\right)\right) \tag{8}$$

$$P\left(lo \mid zh; \theta_{zh\to lo}\right) = \left\{encoder_{zh}, decoder_{lo}\right\} \tag{9}$$

where extractEnc and extractDec are extract functions, which copy the encoder and decoder parameters from the Chinese-Thai model and the Thai-Lao model respectively. Subsequently, the copied parameters are used to initialize the encoder and decoder of the final Chinese-Lao model. $P\left(lo \mid zh; \theta_{zh\to lo}\right)$ is the expected Chinese-Lao NMT model.

As there are significant similarities between Thai and Lao, the distributed representation of the tokens in the two languages will be naturally similar, resulting in a relatively consistent parameter distribution of the Thai and Lao decoders. Actually, we do not have to train a Chinese-Lao NMT model from scratch, but only need to fine-tune the composed Chinese-Thai NMT model obtained above using a small-scale Chinese-Lao parallel corpus. Fortunately, there is an applicable parallel corpus from the ALT, a small-scale but high-quality multilingual dataset that is suitable for the fine-tuning of the final Chinese-Lao translation model. The training process of the Chinese-Lao model based on the maximum likelihood estimation is formulated as:

$$\hat{\theta}_{zh\to lo} = \underset{\theta_{zh\to lo}}{\operatorname{argmax}}\left\{\ell\left(\theta_{zh\to lo}\right)\right\} \tag{10}$$

where lo is the Lao sentences, and the log-likelihood is defined as:

$$\ell\left(\theta_{zh\to lo}\right) = \sum_{n=1}^{N} \log P\left(lo^{(n)} \mid zh^{(n)}; \theta_{zh\to lo}\right) \tag{11}$$

However, before delivering the final model, we must consider a defect that is easily overlooked: the source Chinese encoder is trained to be used by a Thai decoder, while the target Lao decoder is trained to use the outputs of a Thai encoder—not of the source Chinese encoder,

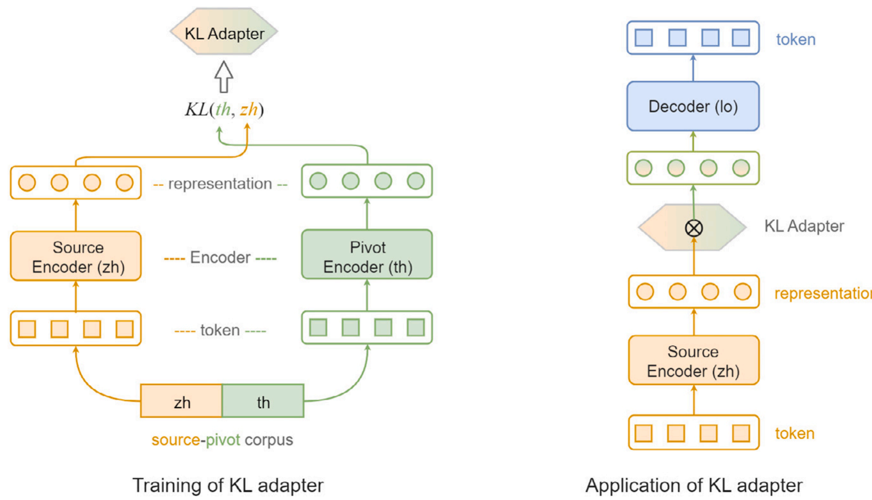

Fig. 4. Training and application of KL adapter. As the KL adapter aims adapting the source encoder outputs to the pivot encoder outputs to which the target decoder is more familiar. For training, we learn a linear affine between the source and pivot representation spaces with a source-pivot parallel corpus. For application, the learned affine is multiplied to the encoder output of all positions in the final source-target tuning step.

which will lead to the problem of semantic space mismatch. To tackle the problem, The core steps are:

1. Encode the source sentence and pivot sentence with the source encoder of the pre-trained source-pivot and pivot-target models respectively, generating the representation h_{zh} of source sentence and the representation h_{th} of pivot sentence.

2. Learn an affine to minimize the distance between the representations h_{zh} and h_{th}:

$$Affine_{h_{zh}\rightarrow h_{th}} = argmin(||h_{th} - h_{zh}||^2) \tag{12}$$

For the convenience of implementation, we build a KL adapter (as shown in Fig. 4) to align the semantic space before the fine-tuning. The loss of the KL adapter is defined by optimizing the KL divergence between the probability distribution of the sentence representation h_{zh} and h_{th} generated by Chinese encoder and Thai encoder respectively:

$$\ell(\theta_{zh\rightarrow th}, \theta_{th\rightarrow lo}) = \sum_{x_{zh},x_{th}} \mathrm{KL}(P(h_{th} \mid x_{th}, \theta_{th\rightarrow lo}) \| P(h_{zh} \mid x_{zh}, \theta_{zh\rightarrow th})) \tag{13}$$

where $x_{zh} \in D_{zh\rightarrow th}$ and $x_{zh} \in D_{th\rightarrow lo}$ are the inputs of the Chinese encoder and Thai encoder respectively.

During the final source-target tuning step, the acquired mapping is applied to the encoder outputs across all positions. This mapping enables the source encoder to produce sentence representations that reside within a comparable space as those of the pivot encoder. By using the KL adapter, the source Chinese encoder and target Lao decoder in the final model have achieved maximum adaptation. Thanks to this adaptive encoder–decoder framework, we can obtain an optimal Chinese-Lao translation model with simple fine-tuning. Since the Chinese-Thai model is trained with a larger corpus, the training of the Chinese encoder is sufficient. Apart from this, by using KL adapter (explicit) and linguistic similarity between Thai and Lao (implicit), the Lao decoder is adaptive to the Chinese encoder to the greatest extent. Compared with using a small-scale Chinese-Lao corpus to train directly from scratch, the performance of the fine-tuned NMT model will be more efficient when using the proposed pivot-based framework. To study the impact of different fine-tuning strategies on the model performance, we adopt five different strategies to realize the fine-tuning and reported the differentiated results in the experimental phase.

5. Experiments

5.1. Experimental setup

Data. We conduct a series of experiments on ALT and homemade datasets to evaluate the performance of the proposed method. ALT is a public multilingual parallel dataset for seven languages: English, Indonesian, Japanese, Khmer, Malay, Myanmar, and Vietnamese. For our task, we choose the trilingual Chinese-Thai-Lao subset, which consists of 20,106 sentence triples. We follow the dataset guideline and bin the Thai-Lao portion that comprises 20,106 sentence pairs into three subsets with the ALT-Standard-Split toolkit: 18,088, 1,000, 1,018 sentence pairs for training, validation, and test datasets, respectively. Due to differences in the difficulty of collecting artificial corpus, the homemade dataset consists of two parts: (1) A 50K Chinese-Thai bilingual parallel corpus which is used for pivoting model training. (2) The 500 sentences Chinese-Lao bilingual parallel corpus which is used for testing the final Chinese-Lao NMT model. This dataset was created using the pipeline manner: network crawling – > machine translation – > manual correction, in which the machine translation stage is implemented by using the Google translate interface.

To maximize the utilization of the above data, we build a composed parallel corpus by combining 18,088 Chinese-Thai sentence pairs selected from the ALT dataset and 50K sentence pairs selected from our homemade dataset to train the Chinese-Thai translation model. While for the training of the Thai-Lao translation model, we first combine the Thai sentences in the homemade corpus and the corresponding Lao sentences (which are translated from Thai sentences in a literal manner with a 10K Thai-Lao parallel lexicon). Then, we integrate it with the 18,088 parallel sentences from the Thai-Lao ALT subset. For the fine-tuning of the Chinese-Lao translation model, we select the Chinese-Lao portion of the ALT dataset as the training data. Details of the training data are presented in Table 2. We pre-process the corpus before feeding it into the translation model. Specifically, Thai sentences are tokenized using the Pythaipiece toolkit,[2] which is based on sentence piece. For Lao tokenization, we apply tokenization using the LaoWordSegmentation toolkit.[3] For Chinese, we apply tokenization using the Jieba toolkit.[4]

Baseline. We thoroughly compare our approach with the following representative baselines:

[2] https://github.com/wannaphong/thai-word-segmentation-sentencepiece.
[3] https://github.com/djkhz/LaoWordSegmentation.
[4] https://github.com/fxsjy/jieba.

Table 2
Experimental data.

Language pair	Parallel data		Pseudo parallel data
	ALT	Homemade	
Chinese-Thai	18,088	50K	/
Chinese-Lao	18,088	/	50K
Thai-Lao	18,088	/	50K

- Direct source→target: standard source-to-target baseline model using bilingual pairs as training data. For specific implementation, we choose vanilla transformer as the direct source→target model which predicts a target sentence in an autoregressive manner relying on self-attention [34].
- Back translation: The standard back-translation model employs a fundamental translation model to translate target language sentences into source language, generating a pseudo parallel corpus that serves as input for the model's retraining. Note that in practical implementation, we use the pre-trained Thai-Lao translation model to translate the Thai portion of the Chinese-Thai corpus into Lao. Although it is different from the standard back translation process, it ensures the fairness of the experiment in terms of data.
- Vanilla transfer (Zoph et al. [16]): a classic transfer learning model, which first train a high-resource language pair model, then transfer the pre-trained parameters to the low-resource language pair model to initialize and constrain training.
- Dynamic vocabulary (Lakew et al. [20]): a transfer learning based NMT model that transferring knowledge from the source-pivot model to the pivot-target model by means of a shared dynamic vocabulary.
- Step-wise pre-training (Kim et al. [9]): pivot-based transfer model, which conducts pre-training with careful parameter freezing.
- Mapping adapter (Kim et al. [9]): pivot-based transfer model, inducing a mapping adapter component to familiarize the pre-trained decoder with the outputs of the pre-trained encoder.
- RL pivot (Tokarchuk, E. et al. [13]): a pivot-based NMT system with the reinforcement learning approach. To enable training on source→target data, it utilizes a non-autoregressive transformer and present an end-to-end pivot-based integrated model.
- LC transfer(Jiang et al. [22]): An approach which applies the transfer learning techniques for the lexical constraint model by elaborately selecting beam search algorithm for lexical constraint measure, and investigates the proper way for transferring parameters across two machine translation models.
- ConsistTL (Li et al. [35]): A transfer learning based NMT method which can continuously transfer knowledge from the parent model during the training of the child model. For each instance in the training of child model, ConsistTL generates a semantically equivalent counterpart for the parent model and promotes consistency in predictions between the two models for that specific instance.
- SEALION (Ong et al. [36]): A large language model (LLM) developed by AI Singapore that better understands Southeast Asia's diverse contexts, languages, and cultures (SEA). We choose its Version 3, a 9B parameter model, with 200 billion tokens from 11+2 Southeast Asian languages as Chinese-Lao translation baseline.
- SEALLM (Nguyen et al. [37]): A LLM that specifically focuses on Southeast Asian (SEA) languages. SeaLLMs are built upon popular English-centric models through continued pre-training with an extended vocabulary, specialized instruction and alignment tuning to better capture the intricacies of regional languages.

Evaluation Metrics. For translation quality evaluation, we adopt BLEU [38] as the evaluation metric and choose the insensitive 4-gram multi-bleu.perl[5] as the calculation script. Apart from this, we use both automatic and human evaluations to measure the fluency of our method. For automatic evaluation, we use RIBES [39] for fluency evaluation. For human evaluation, we follow Snover et al. [40] and ask five professional translators to score adequacy and fluency for each translation, ranging from 1 to 5. Specifically, for fluency evaluation, a score of 1 to 5 indicates the fluency of the translation: 5 = flawless, 4 = good, 3 = non-native, 2 = disfluent, and 1 = incomprehensible. We also conduct adequacy evaluation in the experimental section. Similar to human fluency evaluation, we invite five professional translators to consider how much of the meaning expressed in the reference translation is also expressed in the final translation: 5 = all, 4 = most, 3 = much, 2 = little, and 1= none.

Implement Detail. As the research focuses on low-resource scenario, we follow the low-resource scenario guidelines of Sennrich et al. [41] and adopt prudent parameter settings for both our proposed model and the baseline model. Specifically, we use a self-attention based 2-layer encoder–decoder architecture, in which the number of self-attention heads is set to 4. We also adopt prudent parameter setting and set the word embedding and hidden states to 256. The dropout is set to 0.1 for Chinese-Thai model training and 0.05 for Chinese-Lao model fine-tuning. We use the Adam optimizer [42] for model optimization. The models are evaluated every 1000 steps. All the models follow the Transformer base architecture implemented on an open-source NMT system Thumt [43].

5.2. Experimental results

Translation Quality. Table 3 reports the BLEU evaluation results of the Chinese-Lao translation task. Our model (+KL adapter) achieves conspicuous performance with a p-value < 0.05, as tested by bootstrap resampling [44], outperforming both homogeneous and heterogeneous baselines on all evaluation metrics. In addition, after adopting fine-tuning, the translation performance is further improved. Note that although we introduce an additional 50K of pseudo parallel corpus for the training of the Thai-Lao translation model in our approach, the scale of the parallel corpus used for the training of the Chinese-Lao model is consistent with the baseline system (using the ALT Chinese-Lao subset of 20106 sentence pairs). Therefore, the above evaluation of the Chinese-Lao translation is data-equitable. The quantitative analysis experiment shows that for the Chinese-Lao translation task, using Thai, a language similar to Lao, as the pivot language, and using fine-tuning technology to train the model can significantly improve the quality of the translation. We also selected two large models, SEALION and SEALLM, as baseline systems. Although the translation performance of the large model is not outstanding from the perspective of accuracy, we have observed some interesting phenomena in its translation process, and the relevant analysis is presented in the case study section.

To verify whether the performance improvement is due to the selection of Thai as the pivot language, and whether the similarities between Thai and Lao induce more accurate semantic information transferability, we choose English as the pivot and conduct experiments on the same model setting. Since the ALT dataset is a multilingual parallel dataset, we can achieve a fair comparison based on the same source language (Chinese) and data scale. The corpus for training the Chinese-English model is from the ALT Chinese-English subset combined with the 50K Chinese-English corpus from the IWSLT15 dataset. We adopt two strategies for generating an English-Lao pseudo-parallel corpus: literal translation, where the dictionary is converted from the Thai-Lao lexicon by PanLex,[6] using the Google interface to generate sentence-level translation directly.

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl.
[6] https://panlex.org/.

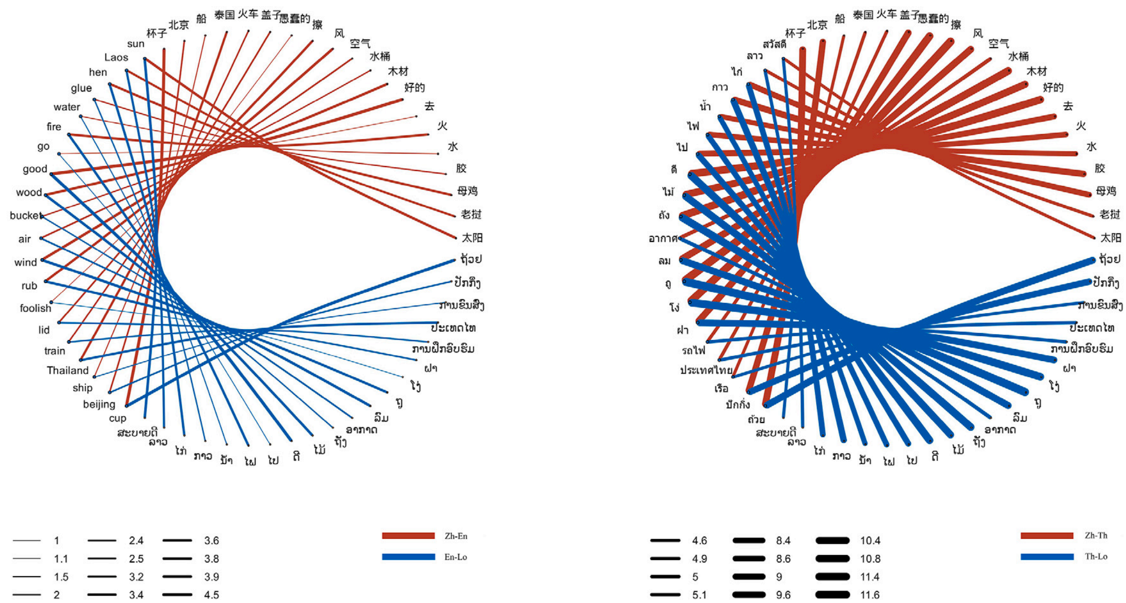

Fig. 5. The semantic transfer effect when using different pivots (Chinese or Thailand), where the width of the line represents the probability of the source language word being correctly translated into the target language.

Table 3
Evaluation results of Chinese-Lao translation task. Transformer is trained with the original ALT Chinese-Lao subset. For our approach, "plain" indicates that initialize the Chinese→Lao model with the source encoder from the pre-trained Chinese→Thai model and the target decoder from the pre-trained pivot→target model, not introducing KL adapters and fine-tuning operations. "parallel" indicates only using the raw parallel corpus (marked as ⟨1⟩ in the Fig. 3) for fine-tuning, while "mixed" indicates both using the raw parallel and synthetic corpus (marked as ⟨2⟩ in the Fig. 3) for fine-tuning. "ALT" and "Homemade" denote the names of the test sets we adopted.

	ALT		Homemade	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Transformer	6.22	77.11	6.15	77.20
Back translation	10.28	73.06	10.10	73.29
Vanilla transfer	11.64	71.65	11.47	71.77
Dynamic vocabulary	12.71	70.94	12.45	71.25
SEALION	12.85	70.84	12.73	71.16
SEALLM	12.80	70.89	12.53	71.22
Step-wise pre-training	13.38	70.79	13.23	70.91
Mapping adapter	13.30	70.71	13.14	70.79
RL pivot	13.59	70.55	13.38	70.67
LC transfer	13.63	70.49	13.34	70.71
ConsistTL	13.77	70.41	13.26	70.78
Our approach (plain)	12.27	71.16	12.29	71.17
+KL adapter	13.14	70.82	12.89	71.01
+fine-tuning (parallel)	13.46	70.60	13.23	70.82
+KL adapter	13.79	70.44	13.55	70.60
+fine-tuning (mixed)	13.81	70.45	13.62	70.55
+KL adapter	**14.06**	**70.22**	**13.78**	**70.06**

Table 4
Performance of our model when applying different pivot languages.

Pivot language	Pseudo-parallel corpus generation manner	BLEU	
		ALT	Homemade
English	literal	9.45	9.17
	sentence-level translation	10.91	10.73
Thai	literal	14.06	13.78

Table 4 shows the different performances of our model when English and Thai are treated separately as pivot languages. When choosing Thai as a pivot, the model achieves a significant improvement compared with choosing English as the pivot language. Apart from this, we can observe that when selecting English as a pivot language, literal translation underperforms sentence-level translation. A possible reason is that the literal translation relies on syntactic and lexical consistency, and it is not suitable for the distant language pair English-Lao. In addition, as shown in Fig. 5, we also used topology visualization techniques [45] to study the translation effect when using different pivot points. We calculate the frequency of translating source language words into target language words using different pivot points, and the width of the lines represents the probability of the source language words being translated into the correct target language words. It can be seen that semantic transmission is more efficient when using Thai as the pivot.

Table 5
Fluency and adequacy evaluation.

	Fluency		Adequacy
	RIBES	Manual	
Transformer	67.35	3.23	3.17
Back translation	68.31	3.34	3.21
Vanilla transfer	68.72	3.40	3.23
Dynamic vocabulary	69.44	3.43	3.27
Step-wise pre-training	70.65	3.49	3.29
Mapping adapter	70.80	3.50	3.31
RL pivot	70.91	3.49	3.38
LC transfer	70.88	3.53	3.36
Our approach (plain)	69.27	3.42	3.26
+KL adapter	69.63	3.44	3.28
+fine-tuning (parallel)	70.45	3.51	3.32
+KL adapter	70.83	3.56	3.35
+fine-tuning (mixed)	70.93	3.55	3.40
+KL adapter	**71.21**	**3.62**	**3.45**

Fluency and Adequacy. To investigate the fluency and adequacy of translation, we evaluate the experimental models on the ALT dataset by automatic and manual evaluation respectively. As shown in Table 5, for automatic fluency evaluation (RIBES), our approach achieves significant improvements compared with baseline models. For manual fluency evaluation, our approach outperforms the baseline system by 0.12–0.39 points. For manual adequacy evaluation, our method also outperforms homogeneous baseline system. Former results demonstrate that our method not only improves the translation effect, but also improves the readability of the translation.

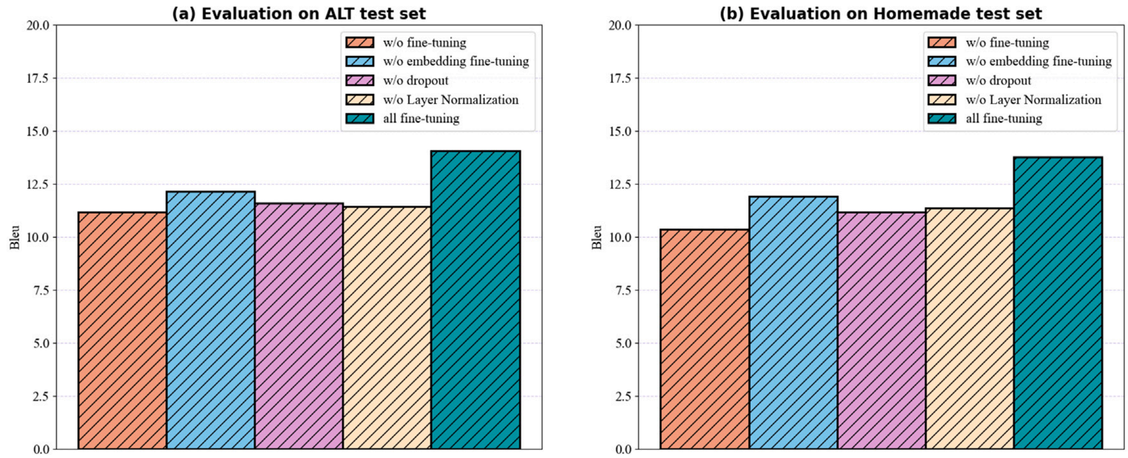

Fig. 6. Ablation study for adopting different fine-tuning strategies.

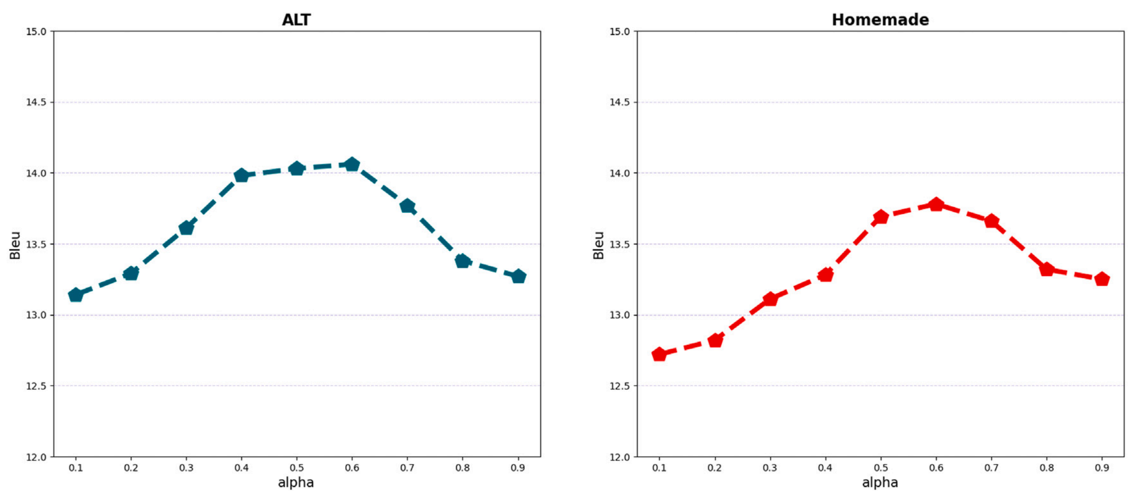

Fig. 7. The influence of hyper-parameter a on the training of Thai-Lao Translation model. Note we conduct the evaluation on the validation set and then validate the conclusion on the test set of ALT and homemade dataset respectively.

Ablation Study w.r.t. different fine-tuning strategies. To investigate the effect of different fine-tuning strategies on the training of the final model, we select three fine-tuning strategies and conduct an ablation study. As shown in Fig. 6, introducing fine-tuning achieves better performance than not using it. In addition, fine-tuning all layers of the model achieves a better performance than fixing the embedding, dropout or normalization layer. The results indicate that the fine-tuning technique is beneficial for our model.

hyper-parameter α. We use a hyper-parameter $\alpha \in [0,1]$ to balance the information proportion during the training of the Thai-Lao Translation model in Section 3.1. We study the influence on the Thai-Lao translation task by raising the value of α. To observe the effect of this hyperparameter, we set the discrete values $\alpha = \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%\}$.

As shown in Fig. 7, for all two test sets, when the α switches from 0.5 to 0.7, model achieves better performance than in the intervals [0.1, 0.4] and [0.8, 0.9]. The results show that we can set the hyper-parameter α in a reasonable interval ([0.5, 0.7]) to maintain the balance between the parallel corpus and the pseudo-parallel corpus.

Case Study. We also study the translation effect of the model using practical examples. As the first example shown in Fig. 8, our model translates the Chinese word "那人" (The man), "处理" (deal with) and "秘密的" (secret) more accurate than baseline model Transformer. A possible reason for this is that by choosing a proper pivot language, our model transfers useful linguistic information to the Lao decoder. The word list in Fig. 10 further verifies this conclusion. For the source Chinese words "那人", "处理" and "秘密的" in the first example, the corresponding words in Thai and Lao are similar in morphology. Moreover, we observed that the corresponding sentences containing these words had similar syntactic structures in the training data. Fig. 9 shows a comparative example with the large model. In the first example, all three models performed well, with the larger model even generating more concise phrases than the reference (see underlined words). The second example illustrates the hallucination problem of LLM, where although the generated translation is very fluent, it is almost unrelated to the source text. Although some words are under-translated in our example, it is still roughly relevant to the original text.

Visualization. We investigate the attention effect by using visualization technology. Since attention mechanism represents the degree of alignment between the source language words and generated target language words in NMT, high-quality alignment will have a positive impact on translation effectiveness. Fig. 11 illustrates the visualization results of the proposed approach for the example in Fig. 8. We find that for the source Chinese words "那人," "处理" and "秘密的," our approach achieves high-quality word alignment and translate them correctly into the correct corresponding words of Lao. Owing to the improvement in word alignment, pivotal words are better reflected in the generated translation, and the fluency of the translation is also improved significantly.

Fig. 8. Case study of Chinese-Lao translation (Transformer baseline). The light coloured font in parentheses denotes the English/Chinese counterpart of the example sentences.

Fig. 9. Case study of Chinese-Lao translation (LLM baselines).

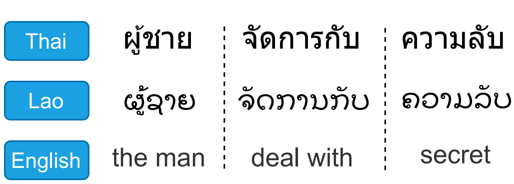

Fig. 10. The morphological similarity of the words in the first sample of Fig. 8.

6. Conclusions

We propose a simple approach for Chinese-Lao translations using a pivot-based strategy. The central idea is to enhance the relationship between the pivot language and target language (Lao). To this end, we choose Thai, which has cross-lingual similarities with Lao, as the pivot language. To integrate the cross-lingual similarities into the domain NMT architecture, we: (1) study the linguistic similarities of the Thai-Lao language pair, (2) propose a simple data augmentation method to create a Thai-Lao pseudo corpus using a simple back-translation method, and (3) propose a novel pivot based translation model for Chinese-Lao translation. Experiments show that our method brings prominent improvements to Chinese-Lao translation compared to strong baseline models.

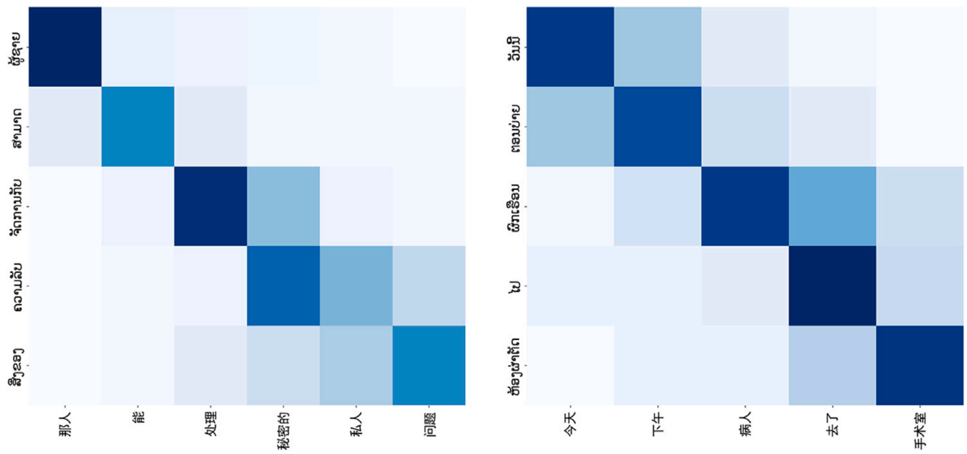

Fig. 11. Attention visualization of the first example in Fig. 8. The horizontal axis represents the input Chinese sentence, while the vertical axis represents the generated Lao sentence. Colour depth indicates varying degrees of attention.

Our approach is also applicable to other low-resource language pairs under the same scenario, such as rich-resource language-Indonesian translations (by choosing Malay as a pivot language, which is highly similar to Indonesian). The limitation of our method is that in zero-resource translation scenarios without source-target corpora, the performance of our method may be affected to some extent due to the limitation of literal translation, but it does not affect the operability of the method. Further work will discuss how to use decoding constraints and other strategies to maintain the performance of the method while excluding literal translation.

CRediT authorship contribution statement

Enchang Zhu: Writing – original draft, Software, Methodology, Conceptualization. **Yuxin Huang:** Writing – review & editing, Supervision, Investigation. **Yantuan Xian:** Supervision, Data curation. **Junguo Zhu:** Visualization, Formal analysis, Data curation. **Minghu Gao:** Visualization, Formal analysis, Data curation. **Zhiqiang Yu:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62241604), Fundamental Research Project of Yunnan Province, China (Grant No. 202301AT070015), Xingdian Talent Support Program Project, China (Grant No. XDYC-QNRC-2023-0388) and Yunnan Province International Joint Laboratory for Intelligent Integration and Application of Ethnic Multilingualism, China (Grant No. 202403AP140014). We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

References

[1] A. Srithirath, P. Seresangtakul, A hybrid approach to lao word segmentation using longest syllable level matching with named entities recognition, in: Proceedings of the 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, IEEE, 2013, pp. 1–5, http://dx.doi.org/10.1109/ECTICon.2013.6559585.

[2] M. Yang, L. Zhou, Z. Yu, S. Gao, J. Guo, Lao named entity recognition based on conditional random fields with simple heuristic information, in: Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD, IEEE, 2015, pp. 1426–1431, http://dx.doi.org/10.1109/FSKD.2015.7382153.

[3] B. Haddow, R. Bawden, A.V.M. Barone, J. Helcl, A. Birch, Survey of low-resource machine translation, Comput. Linguist. 48 (3) (2022) 673–732, http://dx.doi.org/10.1162/coli_a_00446.

[4] A. Chakrabarty, R. Dabre, C. Ding, M. Utiyama, E. Sumita, Low-resource multilingual neural translation using linguistic feature-based relevance mechanisms, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22 (7) (2023) 1–36, http://dx.doi.org/10.1145/3594631.

[5] A.M. Gezmu, A. Nürnberger, Morpheme-based neural machine translation models for low-resource fusion languages, ACM Trans. Asian Low-Resource Lang. Information Process. 22 (9) (2023) 1–19, http://dx.doi.org/10.1145/3610773.

[6] V. Goyal, S. Kumar, D.M. Sharma, Efficient neural machine translation for low-resource languages via exploiting related languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2020, pp. 162–168, http://dx.doi.org/10.18653/v1/2020.acl-srw.22.

[7] Y. Cheng, Y. Cheng, Joint training for pivot-based neural machine translation, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2019, pp. 3974–3980, http://dx.doi.org/10.24963/ijcai.2017/555.

[8] Y. Leng, X. Tan, T. Qin, X.-Y. Li, T.-Y. Liu, Unsupervised pivot translation for distant languages, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 175–183, http://dx.doi.org/10.18653/v1/P19-1017.

[9] Y. Kim, P. Petrov, P. Petrushkov, S. Khadivi, H. Ney, 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, EMNLP-IJCNLP, 2019, pp. 866–876, http://dx.doi.org/10.18653/v1/D19-1080.

[10] Z. Li, H. Zhao, R. Wang, M. Utiyama, E. Sumita, Reference language based unsupervised neural machine translation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 4151–4162, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.371.

[11] R. Dabre, A. Imankulova, M. Kaneko, A. Chakrabarty, Simultaneous multi-pivot neural machine translation, 2021, arXiv preprint arXiv:2104.07410.

[12] G. Chen, S. Ma, Y. Chen, L. Dong, D. Zhang, J. Pan, W. Wang, F. Wei, Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 15–26, http://dx.doi.org/10.18653/v1/2021.emnlp-main.2.

[13] E. Tokarchuk, J. Rosendahl, W. Wang, P. Petrushkov, T. Lancewicki, S. Khadivi, H. Ney, Towards reinforcement learning for pivot-based neural machine translation with non-autoregressive transformer, 2021, arXiv preprint arXiv:2109.13097.

[14] Y. Cai, Y. Cao, X. Wan, Revisiting pivot-based paraphrase generation: Language is not the only optional pivot, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 4255–4268.

[15] T. Zhang, H. Huang, C. Feng, X. Wei, Similarity-aware neural machine translation: reducing human translator efforts by leveraging high-potential sentences with translation memory, Neural Comput. Appl. 32 (23) (2020) 17623–17635, http://dx.doi.org/10.1007/s00521-020-04939-y.

[16] B. Zoph, D. Yuret, J. May, K. Knight, Transfer learning for low-resource neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1568–1575, http://dx.doi.org/10.18653/v1/D16-1163.

[17] M. Zhang, L. Li, Q. Liu, Triangular transfer: Freezing the pivot for triangular machine translation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 644–650, http://dx.doi.org/10.18653/v1/2022.acl-short.72.

[18] T.Q. Nguyen, D. Chiang, Transfer learning across low-resource, related languages for neural machine translation, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2017, pp. 296–301.

[19] T. Kocmi, O. Bojar, Trivial transfer learning for low-resource neural machine translation, in: Proceedings of the Third Conference on Machine Translation: Research Papers, 2018, pp. 244–252, http://dx.doi.org/10.18653/v1/W18-6325.

[20] S.M. Lakew, A. Erofeeva, M. Negri, M. Federico, M. Turchi, Transfer learning in multilingual neural machine translation with dynamic vocabulary, in: Proceedings of the 15th International Conference on Spoken Language Translation, 2018, pp. 54–61.

[21] Y. Kim, Y. Gao, H. Ney, Effective cross-lingual transfer of neural machine translation models without shared vocabularies, 2019, http://dx.doi.org/10.18653/v1/P19-1120, arXiv preprint arXiv:1905.05475.

[22] H. Jiang, C. Zhang, Z. Xin, X. Huang, C. Li, Y. Tai, Transfer learning based on lexical constraint mechanism in low-resource machine translation, Comput. Electr. Eng. 100 (2022) 107856, http://dx.doi.org/10.1016/j.compeleceng.2022.107856.

[23] S. Huang, J. Guo, Z. Yu, Y. Wen, Effective domain awareness and adaptation approach via mask substructure for multi-domain neural machine translation, Neural Comput. Appl. 35 (19) (2023) 14047–14060, http://dx.doi.org/10.1007/s00521-023-08377-4.

[24] S. Liu, X. Liu, D.F. Wong, Z. Li, W. Jiao, L.S. Chao, M. Zhang, kNN-TL: k-nearest-neighbor transfer learning for low-resource neural machine translation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1878–1891.

[25] M. Maimaiti, Y. Liu, H. Luan, M. Sun, Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation, Tsinghua Sci. Technol. 27 (1) (2022) 150–163, http://dx.doi.org/10.26599/TST.2020.9010029.

[26] J. Dong, Transfer learning-based neural machine translation for low-resource languages, ACM Trans. Asian Low-Resour. Lang. Inf. Process. (ISSN: 2375-4699) (2023) http://dx.doi.org/10.1145/3618111.

[27] C. Ding, M. Utiyama, E. Sumita, Similar southeast asian languages: Corpus-based case study on thai-laotian and malay-indonesian, in: Proceedings of the 3rd Workshop on Asian Translation, WAT2016, 2016, pp. 149–156.

[28] K. Singvongsa, P. Seresangtakul, Lao-thai machine translation using statistical model, in: Proceedings of the 13th International Joint Conference on Computer Science and Software Engineering, JCSSE, IEEE, 2016, pp. 1–5.

[29] F.J. Och, H. Ney, A systematic comparison of various statistical alignment models, Comput. Linguist. 29 (1) (2003) 19–51, http://dx.doi.org/10.1162/089120103321337421.

[30] H. Riza, M. Purwoadi, T. Uliniansyah, A.A. Ti, S.M. Aljunied, L.C. Mai, V.T. Thang, N.P. Thai, V. Chea, S. Sam, et al., Introduction of the asian language treebank, in: 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques, O-COCOSDA, IEEE, 2016, pp. 1–6.

[31] H. Isozaki, K. Sudoh, H. Tsukada, K. Duh, HPSG-based preprocessing for english-to-Japanese translation, ACM Trans. Asian Lang. Inf. Process. (TALIP) 11 (3) (2012) 1–16, http://dx.doi.org/10.1145/2334801.2334802.

[32] K. Chen, R. Wang, M. Utiyama, E. Sumita, Neural machine translation with reordering embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1787–1799, http://dx.doi.org/10.18653/v1/P19-1174.

[33] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS '14, MIT Press, Cambridge, MA, USA, 2014, pp. 3104–3112.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[35] Z. Li, X. Liu, D.F. Wong, L.S. Chao, M. Zhang, ConsistTL: Modeling consistency in transfer learning for low-resource neural machine translation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 8383–8394.

[36] D. Ong, P. Limkonchotiwat, SEA-LION (southeast Asian languages in one network): A family of southeast Asian language models, in: Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software, NLP-OSS 2023, Association for Computational Linguistics, Singapore, 2023, p. 245.

[37] X.-P. Nguyen, W. Zhang, X. Li, M. Aljunied, Z. Hu, C. Shen, Y.K. Chia, X. Li, J. Wang, Q. Tan, L. Cheng, G. Chen, Y. Deng, S. Yang, C. Liu, H. Zhang, L. Bing, SeaLLMs - Large language models for southeast Asia, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 294–304.

[38] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[39] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, H. Tsukada, Automatic evaluation of translation quality for distant language pairs, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 944–952.

[40] M. Snover, N. Madnani, B. Dorr, R. Schwartz, Fluency, adequacy, or hter? Exploring different human judgments with a tunable MT metric, in: Proceedings of the Fourth Workshop on Statistical Machine Translation, 2009, pp. 259–268.

[41] R. Sennrich, B. Zhang, Revisiting low-resource neural machine translation: A case study, 2019, arXiv preprint arXiv:1905.11901.

[42] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[43] J. Zhang, Y. Ding, S. Shen, Y. Cheng, M. Sun, H. Luan, Y. Liu, Thumt: An open source toolkit for neural machine translation, 2017, arXiv preprint arXiv:1706.06415.

[44] P. Koehn, Statistical significance tests for machine translation evaluation, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp. 388–395.

[45] M. Hajij, M. Papillon, F. Frantzen, J. Agerberg, I. AlJabea, R. Ballester, C. Battiloro, G. Bernárdez, T. Birdal, A. Brent, et al., Topox: a suite of Python packages for machine learning on topological domains, 2024, arXiv preprint arXiv:2402.02441.