Journal of King Saud University – Computer and Information Sciences 34 (2022) 6552–6563

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

TLSPG: Transfer learning-based semi-supervised pseudo-corpus generation approach for zero-shot translation

Amit Kumar, Rajesh Kumar Mundotiya *, Ajay Pratap, Anil Kumar Singh

Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi 221005, Uttar Pradesh, India

ARTICLE INFO

Article history:
Received 17 September 2021
Revised 15 February 2022
Accepted 7 March 2022
Available online 25 March 2022

Keywords:
Machine Translation
Zero-shot Translation
Transfer Learning
Semi-supervised

ABSTRACT

Machine Translation *(MT)* has come a long way in recent years, but it still suffers from data scarcity issue due to lack of parallel corpora for low (or sometimes zero) resource languages. However, Transfer Learning *(TL)* is one of the directions widely used for low-resource machine translation systems to overcome this issue. Creating parallel corpus for such languages is another way of dealing with data scarcity, yet costly, time-consuming and laborious task. In order to avoid the above listed limitations of parallel corpus formation, we present a TL-based Semi-supervised Pseudo-corpus Generation *(TLSPG)* approach for zero-shot MT systems. It generates the pseudo corpus by exploiting the relatedness between low resource language pairs and zero-resource language pairs via TL approach. It is further empirically ascertained in our experiments that such relatedness helps improve the performance of zero-shot MT systems. Experiments on zero-resource language pairs show that our approach effectively outperforms the existing state-of-the-art models, yielding improvement of $+15.56, +8.13, +3.98$ and $+2$ BLEU points for Bhojpuri→Hindi, Magahi→Hindi, Hindi→Bhojpuri and Hindi→Magahi, respectively.

1. Introduction

Machine Translation *(MT)*, an automatic translation system for conversion of one language into another, gains worldwide attention in the Natural Language Processing *(NLP)* research communities for its contributions (Sutskever et al., 2014). Statistical Machine Translation *(SMT)* and Neural Machine Translation *(NMT)* are two most widely used architectures by MT for language translation. Unlike traditional MT systems (Abercrombie, 2016; Hurskainen and Tiedemann, 2017), SMT is a log-linear framework consisting of language and translation models (Koehn, 2009), whereas NMT is an end-to-end neural network-based encoder-decoder model that predicts the likelihood of a sequence of words using a probabilistic approach. The encoders generate context vectors for input sentences and decoders decode these vectors to generate target sequences. Bahdanau et al. (2014) introduced an attention mechanism in encoder for putting more weights on the words that contained better context vectors of sentences (Bahdanau et al., 2014). Based on the attention mechanism, many improvements have been introduced in NMT, such as Transformer (Vaswani et al., 2017), BART (Lewis et al., 2020) and mBART (Liu et al., 2020) in the recent years.

Both MT models require a huge parallel corpus. NMT has achieved success in dealing with the need of huge parallel resources by introducing various techniques such as back-translation (Edunov et al., 2018), domain adaptation (Chu and Wang, 2018), and fine-tuning (Dabre et al., 2019). NMT covers many scopes of translation for High Resource Languages *(HRLs)* and Low Resource Languages *(LRLs)*. Here, HRLs are the language pairs available in huge amounts to train the model (e.g., German↔English and French↔English). In comparison, LRLs are the language pairs in which training data is insufficient for better learning the context between sentences (e.g., Nepali↔Hindi, Marathi↔Hindi). Insufficient training data works as obstacles for NMT in improving the LRLs' translation quality, leading to context-missing and rare word problems. Some techniques introduced by Sennrich et al. (2016) and Fei et al. (2021) handled such issues.

The problems faced by MTs become more challenging when the availability of training data is almost none or zero. We call such kinds of issues a Zero-Resource Problem *(ZRP)* as described in Fig. 1. Techniques to translate such zero-resource language pairs are known as Zero-Shot Translation *(ZST)*. Some examples of zero-resource language pairs are Magahi↔Hindi, Bhojpuri↔Hindi and Russian↔Hindi (Ojha et al., 2020). Creating parallel corpora for such languages is time-consuming and expensive process due

* Corresponding author.
E-mail addresses: amitkumar.rs.cse17@iitbhu.ac.in (A. Kumar), rajeshkm.rs.cse16@iitbhu.ac.in (R.K. Mundotiya), ajay.cse@iitbhu.ac.in (A. Pratap), aksingh.cse@iitbhu.ac.in (A.K. Singh).

https://doi.org/10.1016/j.jksuci.2022.03.008
1319-1578/

to manual involvement of many language experts. Several works have been done on the NMT that support the translation of zero-resource language pairs- e.g., multilingual and pivot-based translations (Firat et al., 2016; Johnson et al., 2017; Lu et al., 2018; Liu et al., 2020). Multilingual models usually generalize in a better way due to inclusion of multiple languages (Dabre et al., 2020). However, sometimes this is not valid for morphologically rich languages due to differences in morphological complexity. Pivot-based MT is also one of the traditional approaches for producing translation of zero-resource language pairs. However, training the model via pivot-based approach leads to fluency issues (Nasution et al., 2017). To address above listed problems of ZST, we propose a Transfer Learning-based Semi-supervised Pseudo-corpus Generation *(TLSPG)* approach for translation of zero-resource languages that uses semi-supervised learning to exploit similarities between low and zero-resource language pairs.

The proposed TLSPG approach is motivated by the work of Kumar et al. (2020) built on the hybrid architecture of SMT and NMT. TLSPG generates the pseudo corpus by leveraging the relatedness between low and zero-resource language pairs and learns the context of sentences in a semi-supervised way using Transfer Learning *(TL)*. Unlike multiple HRLs and LRLs in multilingual-based ZST, we use only a single LRLs parallel corpus to develop an MT system for zero resource languages. We demonstrate the experiments on Nepali (NE)↔Hindi (HI), Bhojpuri (BHO)↔Hindi (HI) and Magahi (MAG)↔Hindi (HI) language pairs. In our experiments, Nepali↔Hindi is used to generate the zero-resource language pairs (Bhojpuri↔Hindi and Magahi↔Hindi) by leveraging their relatedness via TL. All the demonstrated languages are mainly spoken in South Asian countries. Applications of ZST can be supported in different fields such as smart healthcare (Mutal et al., 2019; Skianis et al., 2020), military and defence (Klavans et al., 2018), finance (Ghaddar and Langlais, 2020) and e-commerce (Calixto et al., 2017). For instance, use of developed model can be helpful in removing the communication barrier between medical practitioners and local language speakers in a healthcare domain (Stickland et al., 2021).

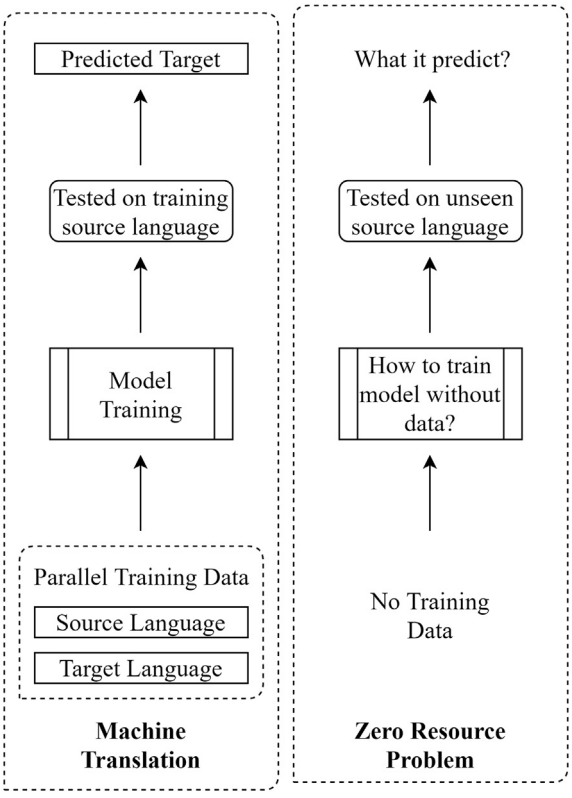

Fig. 1. Zero Resource Problem in MT.

Table 1
Relatedness features between languages.

Languages	Language Family	Script	Word Order
Bhojpuri	Indo-Aryan	Devanagari	S-O-V
Hindi	Indo-Aryan	Devanagari	S-O-V
Magahi	Indo-Aryan	Devanagari	S-O-V
Nepali	Indo-Aryan	Devanagari	S-O-V

Note:- S: Subject, O: Object, V: Verb.

Table 2
Sentence examples.

Languages	Sentences
Bhojpuri	u Apana celA ke xaraxa buJa gailana.
Hindi	Axi meM parameSvara ne AkASa Ora pqWvI kI sqRti kI.
Magahi	jirI sA bolalUz wo alabala bake lagalA.
Nepali	gretara noedA vestako kAlo bAxala Gatne nAma liiraheko CEna.

Note:- All the languages are represented via WX (Diwakar et al., 2010).

Based on sharing common characteristics described in Table 1, Bhojpuri, Magahi, Nepali and Hindi are considered related languages. Moreover, sentence examples given in Table 2 demonstrate relatedness between above considered languages based on commonality in writing script, word ordering, and language family.

Specifically, the contributions of this paper are summarized as follows:

- Propose TLSPG approach for ZST to overcome parallel data limitation of existing NMT models.
- Unlike the existing multilingual-based ZST models (Firat et al., 2016; Johnson et al., 2017; Liu et al., 2020), proposed approach leverages the relatedness of single LRLs pair as a semi-supervised TL technique to improve the performance and validates through empirical analysis.
- Moreover, perform statistical significance analysis and measure robustness through training the model on different subsamples of synthetically generated data and comparing with the other state-of-the-art techniques available in the literature.

The rest of the paper is organized as follows: Closely related works are reviewed in Section 2. Problem formulation and proposed model are discussed in Section 3. Experimental data set and setup are given in Section 4. Obtained results and respective analysis are given in Section 5. Finally, Section 6 concludes this work.

2. Related Work

In this section, we closely review the existing ZST systems shown in Table 3. Firat et al. (2016) proposed a finetuning algorithm for multiway, multilingual NMT model to translate zero-resource language pairs (Firat et al., 2016). Johnson et al. (2017) fed all training data into a single NMT engine and trained the model (Johnson et al., 2017). In the work of Sestorain et al. (2018), authors demonstrated a zero-shot system consisting of reinforcement and dual learning.

In the work of Lakew et al. (2018), authors have suggested a multilingual NMT on a zero-shot direction based on monolingual data and demonstrated that the self-learning technique improves the efficiency of multilingual zero-shot directions by using bilingual parallel corpora for training. Lu et al. (2018) demonstrated a multilingual encoder-decoder NMT architecture with an explicit neural interlingua for performing direct ZST (Lu et al., 2018).

Pham et al. (2019) designed a setup by setting a "Chain" of languages for 12 language pairs on the standard IWSLT 2017 multilingual benchmark (Pham et al., 2019). Hokamp et al. (2019) presented a multilingual MT system for ZST on 110 unique translation directions trained on WMT 2019 shared parallel task datasets and evaluated by creating gold sets for zero-shot pairs in TED talks multi-parallel datasets (Hokamp et al., 2019).

Gu et al. (2019) addressed the degeneracy issue by quantitatively analyzing the mutual information between the language of the source and decoded sentences (Gu et al., 2019). Arivazhagan et al. (2019b) performed a zero-shot experiment on 103 languages trained on 25 billion examples (Arivazhagan et al., 2019b). Arivazhagan et al. (2019a) proposed an auxiliary loss forcing the model to learn the source language invariant representations that improve generalization (Arivazhagan et al., 2019a). Al-Shedivat and Parikh (2019) focused on ZST generalization and proposed a consistent agreement-based learning approach for zero-shot translation (Al-Shedivat and Parikh, 2019). Zhang et al. (2020) demonstrated the feasibility of back-translation to allow for massively ZST and conduct the experiments on the multilingual dataset (Zhang et al., 2020).

Kumar et al. (2020) proposed a bilingual based ZST system for Bhojpuri↔Hindi and Magahi↔Hindi language pairs (Kumar et al., 2020). It is based on an unsupervised domain adaptation approach. Liu et al. (2020) presented *mBART* - a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective (Liu et al., 2020). Lakew et al. (2021) proposed a novel zero-shot NMT approach, which includes three stages: initialization, augmentation, and training for constructing a self-learning cycle of zero-shot pair (Lakew et al., 2021).

As discussed above, most of the existing methods for ZST are mainly outcomes of the multilingual NMT model and specifically trained on the combinations of HRLs and LRLs to improve the qualities of the translation of zero-shot languages. The insufficient availability of parallel corpora acts as a hindrance in developing the ZST systems. In comparison with existing methods, our proposed approach based on TL works on leveraging the relatedness of LRLs pairs without any help of HRLs, shows the drastic improvement for zero-shot language pairs.

3. Transfer Learning-based Semi-supervised Pseudo-corpus Generation

This section discusses the framework of our proposed model to handle the ZRP. We propose a framework based on transfer learning and name it as Transfer Learning-based Semi-supervised Pseudo-corpus Generation *(TLSPG)* approach. It consists of three modules: Transformer-based Semi-supervised Learning *(TSL)*, Moses-based Semi-supervised Learning *(MSL)* and TL-based pseudo-corpus generation. TSL and MSL modules pretrain the model for zero-resource language pairs based on a semi-supervised learning. TL-based pseudo-corpus generation module generates the parallel aligned corpus for zero-resource language pairs via pre-trained TSL and MSL modules. Then synthetic parallel corpus is generated by merging the parallel corpus of related language with pseudo-corpus generated data and training the translation model via Transformer or Moses based translation system.

3.1. TSL

TSL is a semi-supervised transfer learning approach based on training the NMT model via transformer architecture. We train the transformer with five number of encoder and decoder stacks. In order to fill the gap of training language pair in ZST, TSL takes zero-resource related language pair as input to train the transformer. Before training, TSL pre-processes the training sentences via sentencepiece unsupervised tokenizer and converts the sentences into subword tokens. To train the model, generated subword tokens are added to positional encoding in the forms of subword embedding and given as input to encoder and decoder layers as shown in Fig. 2. TSL computes the attention in Transformer as follows:

$$attn = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, V and d_k represent query, key, value and dimension of the key generated from input sequences, respectively.

The cross-entropy loss function L_r used to train the transformer-based NMT model is defined below:

Table 3
Comparative overview of the closely related existing models.

Papers	Types of MT		Techniques		Training model	Zero-resource Language pairs
	Bilingual	Multilingual	Finetuning	Pivot	Neural	
Firat et al. (2016)		✓	✓		GRU	ES→FR
Sestorain et al. (2018)		✓			LSTM	ES↔FR, ES↔RU, RU↔FR
Johnson et al. (2017)		✓			LSTM	PT→ES, ES→JA, EN↔{BE, RU, UK}
Lakew et al. (2018)	✓	✓			RNN and Transformer	IT↔RO
Lu et al. (2018)		✓			LSTM	FR↔RU, ES↔ZH, ES↔FR
Pham et al. (2019)		✓			Transformer	EN↔RO, DE↔IT, EN↔NL, NL↔IT, DE↔RO, NL↔RO
Hokamp et al. (2019)		✓			Transformer	CS, DE, FI, GU, KK, LT, RU, TR, ZH, FR (88 directions)
Gu et al. (2019)		✓			Transformer	DE→IT, DE→NL, AR→RU, AR→ZH, RU→ZH
Arivazhagan et al. (2019b)		✓			Transformer	DE→FR, BE→RU, Yi→DE, FR→ZH, HI→FI, RU→FI
Arivazhagan et al. (2019a)		✓	✓	✓	Transformer	DE↔FR
Al-Shedivat and Parikh (2019)		✓	✓		LSTM	ES↔DE, ES↔FR, DE↔FR
Zhang et al. (2020)		✓	✓	✓	Transformer	OPUS-100
Kumar et al. (2020)	✓				Transformer	BHO↔HI, MAG↔HI
Liu et al. (2020)		✓	✓		mBART	NL↔EN, AR↔EN, NL↔DE
Lakew et al. (2021)	✓	✓		✓	Transformer	AZ↔EN, BE↔EN, GL↔EN, SK↔EN

Note: EN: English, ES: Spanish, DE:German, FR: France, RU: Russian, PT: Portuguese, JA: Japanese, ZH: Chinese, GU: Gujarati, HI: Hindi, IT: Italian, RO: Romanian, BHO: Bhojpuri, MAG: Magahi, BE: Belarusian, UK: Ukrainian, CS: Czech, NL: Dutch, FI: Finnish, KK: Kazakh, LT: Lithuanian, TR: Turkish, AR: Arabic, AZ: Azerbaijani, GL: Galician, SK: Slovak, GRU: Gated Recurrent Unit, LSTM: Long Short-Term Memory, RNN: Recurrent Neural Network.

$$L_r = -\sum_{(X_r,Y_r)\in D_r} \hat{p}_{(X_r,Y_r)} \log_{10} P(Y_r|X_r), \tag{2}$$

where X_r and Y_r are source and target sentences belonging to zero-shot related training language pairs D_r, respectively. Moreover, $\hat{p}_{(X_r,Y_r)}$ is the gold distribution of X_r. The softmax function used to convert the predicted subword embeddings into probabilities is defined as follows:

$$p(Y_t) = \frac{\exp(Y_t)}{\sum_{j=1}^{M} \exp(Y_j)}, \tag{3}$$

where, M denotes the total number of unique words known by the model for the generated subword vector Y_t at time step t.

The decoder decodes the predicted target probabilities and passes them to the beam search (Fig. 2). Beam search gives the best predicted target-side subword tokens. Then detokenization is performed on predicted target-side subword tokens, and the model predicts the target sequences. For prediction, we give the zero-resource test sentence as input to the model. Finally, we get the ZST model as an output of the TSL approach.

3.2. MSL

MSL is a semi-supervised TL approach rely on training a phrase-based SMT system via Moses (Koehn et al., 2007) framework. In order to fill the gap of training language pairs in ZST, MSL takes related language pairs as input to train Moses as shown in Fig. 3. Before training, MSL preprocesses the training sentences via Moses tokenizer and limits the sentences up to 80 length. MSL, a Moses-based (log-linear) framework relies on two modules: language and translation. KenLM (Heafield, 2011) trains the language model on the target side monolingual corpus of related language pairs. Translation model consists of phrase translation and distortion probabilities. For translation, MSL uses GIZA++ (Och and Ney, 2003) for training on related language parallel corpus. We train

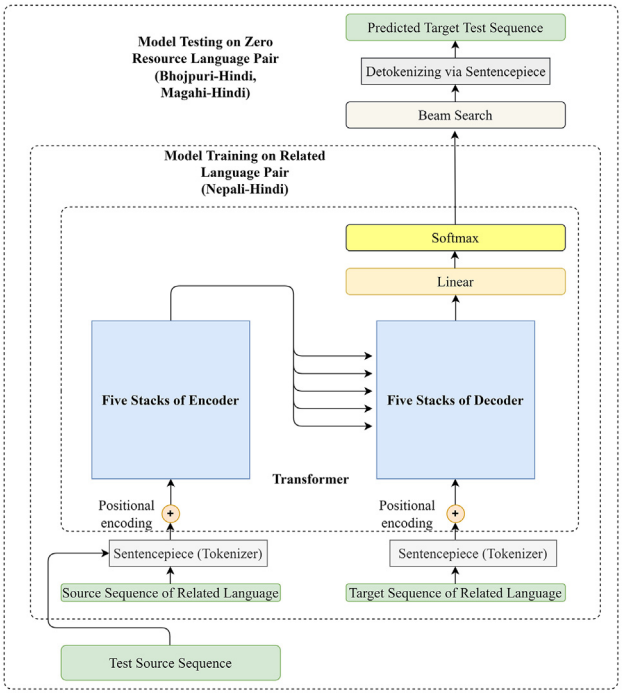

Fig. 2. TSL for zero-resource language pair.

overall MSL on Moses decoder. The decoder decodes the predicted target tokens. Then detokenization is performed on tokenized sequences and final target sequences are predicted. MSL computes the best target translation e_{best} for a source input sentence f as follows:

$$\begin{aligned} e_{best} &= \text{argmax}_e p(e|f) \\ &= \text{argmax}_e p(f|e)\, p_{LM}(e), \end{aligned} \tag{4}$$

where $p_{LM}(e)$ and $p(f|e)$ are language and translation models, respectively.

A phrase-based log-linear MSL model decomposes $p(f|e)$ into $p(\bar{f}_1^I|\bar{e}_1^I)$ as given in the following (Koehn, 2009):

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(start_i - end_{i-1} - 1), \tag{5}$$

where, I is the number of phrases $\bar{f}_i$ broken from f, ϕ is phrase translation probability, $d(.)$ is distortion probability, $start_i$ is the position of the first word of the source input phrase that translates to the i^{th} target phrase and end_i is the position of the last word of that source phrase.

Finally, in order to test the model, we give the zero-resource test data as input for prediction and get Moses-based ZST model as an output.

3.3. TL-based pseudo-corpus generation

In this section, we discuss the pseudo-corpus generation method based on the pre-trained TSL and MSL models. Pseudo-corpus generation module with TLSPG approach is demonstrated in Fig. 4. TLSPG first applies the pretrained TSL or MSL model on target-side monolingual data of zero-resource language pairs to generate the predicted source-side monolingual sentences. Then both the target-side monolingual sentences of zero-resource language pairs and predicted source-side monolingual sentences are aligned parallelly in the direction of Source→Target. TLSPG merges the generated aligned parallel data with Source→Target parallel corpus of related language pairs to create synthetic Source→Target parallel corpus for zero-resource language pairs.

3.4. Model Training

In this section, we discuss the training of the final ZST model. We train the final ZST via the Transformer and the Moses models. For Transformer, we define the cross entropy loss function to train ZST model as follows:

$$L_{ZST} = -\sum_{(X_{syn},Y_{syn})\in D_{syn}} \hat{p}_{(X_{syn},Y_{syn})} \log_{10} P(Y_{syn}|X_{syn}), \tag{6}$$

where, X_{syn} and Y_{syn} represent source and target synthetic generated parallel sentences belonging to synthetic generated training corpus D_{syn}, respectively. Moreover, $\hat{p}_{(X_{syn},Y_{syn})}$ is the gold distribution of X_{syn}.

For Moses, we use the same objective function defined in Eq. (4) for synthetic generated corpus. In order to get final ZST model, we train the translation model on pseudo generated corpus with following variations:

- *A* **Transformer model training on TSL generated corpus architecture:** We train the Transformer on TSL-based generated synthetic corpus with five layers of encoder and decoder to train the Source→Target ZST model.
- *B* **Moses training on MSL generated corpus architecture:** We train the Moses on MSL-based generated synthetic corpus with 6-gram KenLM language model to train the Source→Target ZST model.

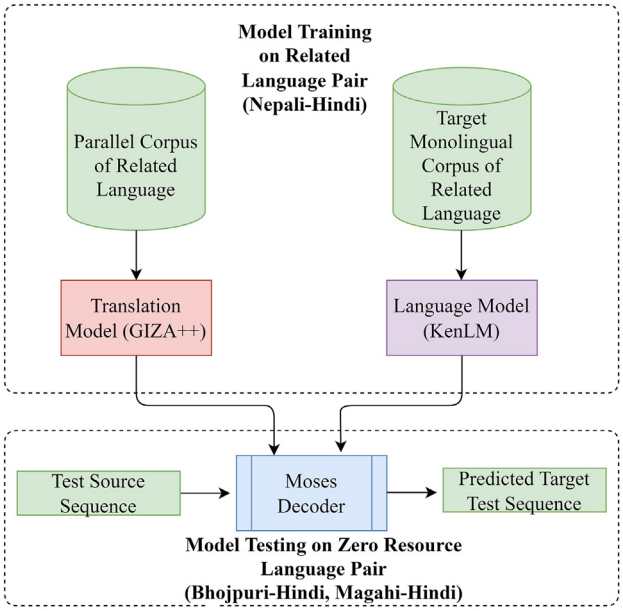

Fig. 3. MSL for zero-resource language pair.

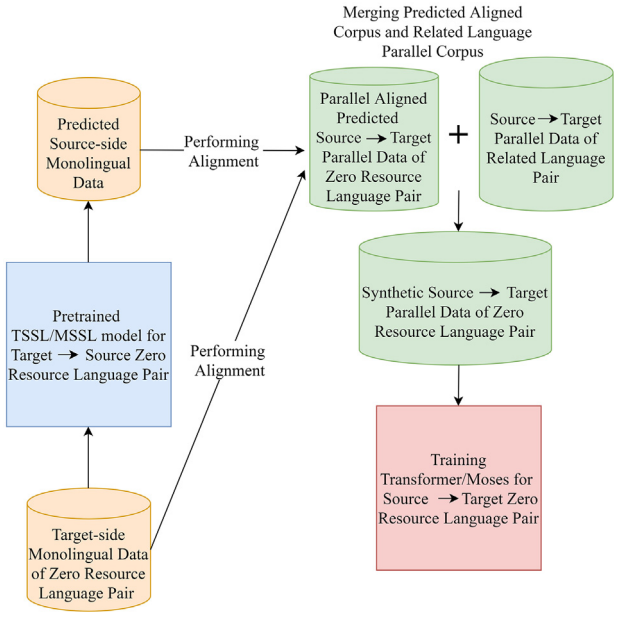

Fig. 4. TLSPG approach.

C **Transformer training on MSL generated corpus architecture:** We train the Transformer on MSL-based generated synthetic corpus with five layers of encoder and decoder to train the Source→Target ZST model.

D **Moses training on TSL generated corpus architecture:** We train the Moses on TSL-based generated synthetic corpus with 6-gram KenLM language model to train the Source→Target ZST model.

4. Data and Experimental Setup

In this section, we discuss the datasets and the experimental settings required to execute the models and analyze the results.

4.1. Data Preparation

We evaluate our proposed model on two language pairs (four translation directions): Hindi→Bhojpuri, Bhojpuri→ Hindi, Hindi→Magahi and Magahi→Hindi. Since, all the used language pairs have zero training data, we employ the model training on Nepali↔Hindi parallel corpus for semi-supervised TL learning. Training and development dataset for Nepali-Hindi parallel corpus are obtained from WMT 2019 similar language shared task (Barrault et al., 2019), Opus (Tiedemann, 2012), and TDIL[1]. In addition, the LoResMT2020[2] shared task provided a monolingual corpus as well as the development and test sets for the Ojha et al. (2020). Table 4 summarises data statistics. All datasets are preprocessed using SentencePiece[3] tokenizer. The proposed model learns 5000 merge operations and restricts the source and target vocabulary to the most frequent 5000 tokens for the Transformer architecture.

4.2. Experimental Setup

This part discusses the experimental settings required to train the TLSPG and baseline models in the following:

[1] http://www.tdil-dc.in/index.php?lang=en
[2] https://sites.google.com/view/loresmt/loresmt-2020
[3] https://github.com/google/sentencepiece

4.2.1. TLSPG

For TSL, we use the NMT model based on Transformer architecture. Transformer has been trained and evaluated on the open-source Fairseq toolkit (Ott et al., 2019). We have trained the model on the default parameters of Kumar et al. (2020) as described in Table 5, for better comparison. For MSL, we use Moses[4], a phrase-based statistical MT model. We apply GIZA++ and KenLM to train the translation and language models of Moses, respectively. Moreover, GIZA++ is employed for phrase alignment based on the Markov model. We also use Mert for minimum error rate training, i.e., to tune the model. We train the KenLM on the different setups of 1 to 6-gram and consider 6-gram in our experiments for a critical analysis of models. In the pseudo-corpus generation module, Transformer and Moses are trained on the same settings as described in TSL and MSL models.

4.2.2. Baselines

We use mBART (Liu et al., 2020), a state-of-the-art multilingual and zero-shot MT approach to compare our proposed TLSPG model. We employ the NE↔HI component of the multilingual NMT method from the pre-trained *mbart.cc*25 (Liu et al., 2020) model to directly evaluate it on BHO↔HI and MAG↔HI test sets in zero-shot conditions due to similarity among NE↔HI, BHO↔HI and MAG↔HI language pairs. In addition to mBART, we also compare our model performance with the work done by Kumar et al. (2020) on the same training and test dataset.

5. Results and Analysis

We evaluate our model on three metrics: BLEU (Papineni et al., 2002), chrF2 (Popović, 2015), and TER (Snover et al., 2006). To compute each metric, we use SacreBLEU (Post, 2018) tool. From the obtained scores of each metric listed in Table 6 on different models, we see that the proposed approach outperforms the existing state-of-the-art models in all three metrics with wider margin. The reported scores of each metric also show a lot of co-relation between each other. We can conclude that the Moses-based system performs better than Transformer. The similarity (relatedness)

[4] http://www.statmt.org/moses/

Table 4
Description of corpus statistics.

Languages	Types	Sentences
NE↔HI**	Training	136991
	Development	3000
HI*	Training	473605
BHO*	Training	91131
MAG*	Training	148606
BHO↔HI**	Development	500
	Test	500
MAG↔HI**	Development	500
	Test	500

* Monolingual data.
** Parallel data.

Table 5
Experimental setup used to train the TSL model.

Parameter	Value
Model	Transformer
Encoder and Decoder layers	5
Encoder embedding dimension	512
Decoder embedding dimension	512
Encoder attention heads	2
Decoder attention heads	2
Dropout	0.4
Attention dropout	0.2
Optimizer	Adam
Learning rate scheduler	inverse sqrt
Learning rate	1e-3
Minimum learning rate	1e-9
Adam-betas	(0.9, 0.98)
Number of epochs	100

factors shared by Bhojpuri, Magahi, Nepali and Hindi account for the Moses-based system's superior performance over a transformer-based approach. Because of their similarities, the source language (Bhojpuri, Magahi and Nepali) and target language (Hindi) in our case follow similar sentence structures. As a result, Moses generates the phrases having the same structure in both languages and performs phrase translations. Therefore, phrase translation enhances performance in a Moses-based system.

5.1. Impact of TSL and MSL

Without using the pseudo-corpus generation approach, TSL and MSL in TLSPG get an improvement of +15 and +30 BLEU on BHO→HI respectively, +9 and +13 BLEU on MAG→HI, respectively, +2 and +1 BLEU on HI→BHO respectively, and +2 and +1 BLEU on HI→MAG respectively compared to mBART model. One of the possible reasons behind improvement is the high relatedness between the language pairs.

5.2. Impact of TLSPG

Our proposed method, TLSPG, gets an improvement of +32.43, +18.34, +6.36 and +4.96 BLEU points for BHO→HI, MAG→HI, HI→BHO and HI→MAG, respectively compared to mBART model. We see that our proposed approach outperforms the state-of-the-art models with a wide margin. This shows that relatedness can play a major role in improving the ZST systems. Apart from these improvements, we also noticed a large variation of BLEU between X→HI and HI→X (where X are BHO and MAG). Such a large variation of BLEU while changing the language direction depends on the complexity of the languages described in the next section.

5.3. Relatedness between languages

In this part, we perform some empirical analysis on the relatedness factor of languages to analyze the large improvement in score. We use a corpus based approach, SSNGLMScore (Mundotiya et al., 2021), to measure the similarity (relatedness) between languages.

5.3.1. Cross-lingual similarity between languages using SSNGLMScore

We use the similarity metric given by Mundotiya et al. (2021), called as SSNGLMScore, to measure the relatedness between the languages, defined as follows:

$$SS_{G,H} = \sum_{H=1}^{n} Score(G(H)), \tag{7}$$

where *SS* stands for Scaled Sum of *n*-gram language model scores.

$$MSS_{G,H} = \frac{SS_{G,H} - \min(SS_{LM,TL})}{\max(SS_{LM,TL}) - \min(SS_{LM,TL})}, \tag{8}$$

where, LM and TL represent language model and test language, respectively. Moreover, $G \in$ LM(Nepali, Bhojpuri, Hindi, Magahi) and n is the total number of sentences in the test language $H \in$ TL (Nepali, Bhojpuri, Hindi, Magahi). We train the G using a 6-gram KenLM model on monolingual corpus described in Table 4. Each language model G is tested on each language H and scores are reported.

Table 7 lists the cross-lingual similarity scores of Bhojpuri, Magahi, Nepali and Hindi with each other. The values in Table 7 indicate how closely languages are related to one another. It

Table 6
Experimental results for different language pairs.

Language Pairs	Scores	mBART	Kumar et al. (2020)	MSL	TSL	A	B	C	D
Bhojpuri→Hindi	BLEU	2.63	19.5	32.78	17.80	19.44	35.06	19.49	32.94
	chrF2	0.46	-	0.53	0.56	0.58	0.57	0.59	0.57
	TER	1.000	-	0.459	0.656	0.609	0.549	0.595	0.565
Magahi→Hindi	BLEU	3.50	13.71	16.67	12.58	14.94	20.54	16.14	21.84
	chrF2	0.43	-	0.44	0.43	0.46	0.43	0.49	0.48
	TER	1.000	-	0.601	0.725	0.662	0.652	0.654	0.618
Hindi→Bhojpuri	BLEU	0.16	2.54	1.16	2.61	3.78	4.77	2.56	6.52
	chrF2	0.08	-	0.16	0.15	0.17	0.24	0.16	0.25
	TER	1.000	-	1.313	0.995	0.971	0.803	0.982	0.780
Hindi→Magahi	BLEU	0.19	3.16	1.17	2.89	3.18	3.50	2.39	5.15
	chrF2	0.07	-	0.16	0.14	0.17	0.22	0.15	0.24
	TER	1.000	-	1.376	1.027	1.023	0.850	1.025	0.896

A: Transformer model training on TSL generated corpus architecture.
B: Moses training on MSL generated corpus architecture.
C: Transformer training on MSL generated corpus architecture.
D: Moses training on TSL generated corpus architecture.

Table 7
SSNGLMScore.

Score	BHO	MAG	NE	HI
BHO	-	0.3015	0.2823	0.2996
MAG	0.3015	-	0.3635	0.3366
NE	0.2823	0.3635	-	0.4845
HI	0.2996	0.3366	0.4845	-

Table 8
Entropy and Type-to-Token Ratio.

Languages	Entropy	Type-to-Token Ratio
Hindi	5.1474	0.0361
Nepali	5.6140	0.1221
Bhojpuri	4.9658	0.0527
Magahi	5.1480	0.0531

empirically justifies the relatedness between languages that show highly co-relation with the results described in Table 6. This relatedness between languages aids the performance of our proposed model for zero-resource languages, as shown in Table 6. We see the performance of Hindi→Bhojpuri is close to Hindi→Magahi. The models for Hindi→Bhojpuri and Hindi→Magahi are built by applying a transfer learning approach on Hindi→Nepali language pair. Moreover, based on Table 8, Nepali is morphologically more complex than Bhojpuri and Magahi. The initial pair of translations for Hindi→Bhojpuri and Hindi→Magahi was Hindi→Nepali, with Nepali as the target language, which decreases MT performance (Mi et al., 2020). The complexity between languages hurts the transferable parameters of Bhojpuri and Magahi. Hence, the BLUE's differences between Hindi→Bhojpuri and Hindi→Magahi are very close. The reason for the better score of Bhojpuri→Hindi than Magahi→Hindi is the out-of-vocabulary difference between

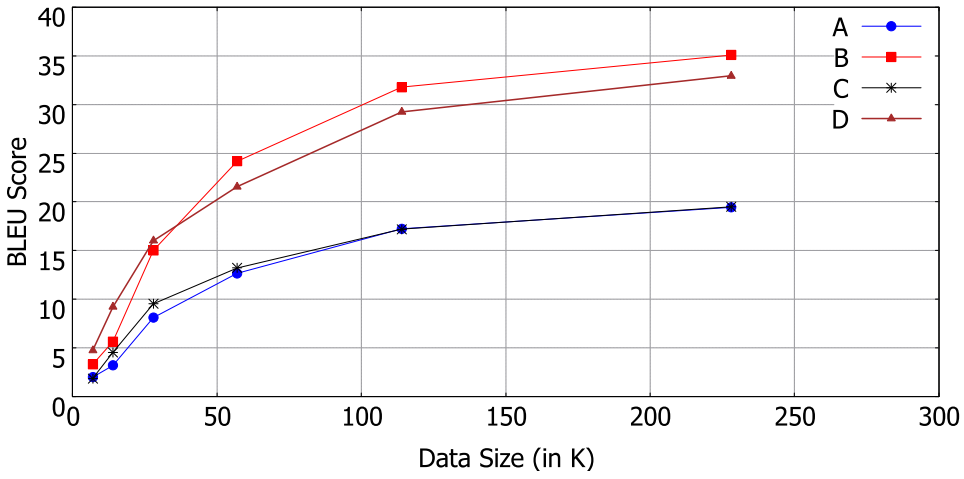

Fig. 5. Comparison between BLEU score and subsamples of training data for Bhojpuri→Hindi.

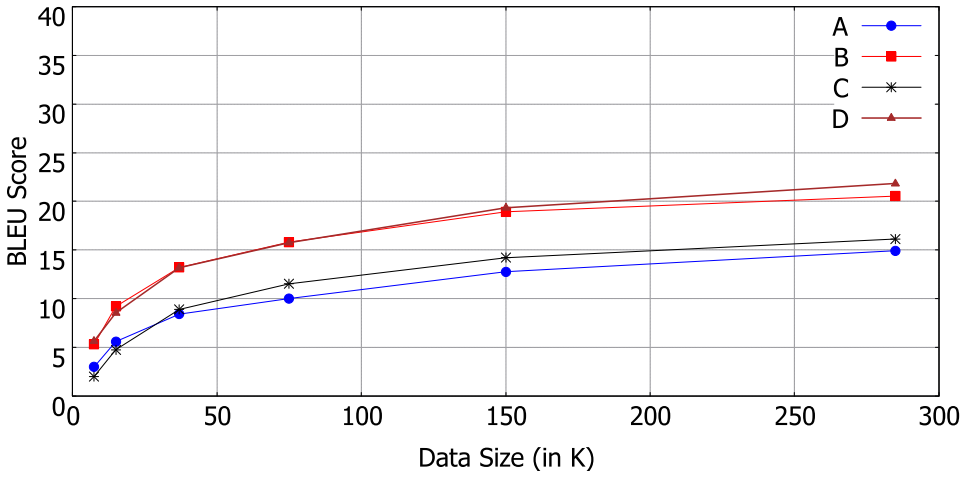

Fig. 6. Comparison between BLEU score and subsamples of training data for Magahi→Hindi.

Table 9
Experiments on using LSTM instead of Transformer.

Language Pairs	Metrics	LSSL	A_{lstm}	C_{lstm}	D_{lstm}
Bhojpuri→Hindi	BLEU	16.7	18.41	17.21	29.43
	chrF2	0.54	0.57	0.56	0.55
	TER	0.661	0.611	0.601	0.569
Magahi→Hindi	BLEU	11.0	12.28	15.10	18.41
	chrF2	0.41	0.46	0.48	0.48
	TER	0.695	0.662	0.657	0.622
Hindi→Bhojpuri	BLEU	2.52	2.91	2.51	5.48
	chrF2	0.12	0.15	0.11	0.19
	TER	1.061	0.975	0.991	0.7889
Hindi→Magahi	BLEU	2.91	2.99	2.28	5.09
	chrF2	0.16	0.16	0.15	0.19
	TER	1.121	1.027	1.029	0.899

A_{LSTM}: LSTM model training on LSSL generated corpus architecture
C_{LSTM}: LSTM training on MSL generated corpus architecture
D_{LSTM}: Moses training on LSSL generated corpus architecture.

Magahi and Bhojpuri with Nepali. Thus, we analyze out-of-vocabulary in the following Section.

5.4. Impact of out-of-vocabulary

Out-of-vocabulary is the collection of words present in test data but absent in training data. Reason for the better score of Bhojpuri→Hindi than Magahi→Hindi is more out-of-vocabulary present in Magahi test data than that of Bhojpuri test data. Since we use Nepali→Hindi as training data because of its relatedness with Bhojpuri→Hindi and Magahi→Hindi language pairs, we compute the out-of-vocabulary ratio between the source languages such as between Nepali and Bhojpuri, and between Nepali and Magahi. For this, we perform the set operations on training data of Nepali with test data of Bhojpuri and Magahi. Out-of-vocabulary ratio of Bhojpuri is 45.79% and Magahi is 69.98% concerning Nepali. We find that Magahi has 3.16 times more out-of-vocabulary words compared to Bhojpuri. Therefore, this is the reason behind the better BLEU score of Bhojpuri→Hindi compared to Magahi→Hindi despite Magahi being more similar to Nepali than Bhojpuri.

5.5. Impact of language complexity

Our studies primarily include morphologically diverse languages. To correlate our findings with the morphological richness of languages, we have used corpus-based complexity scales.

5.5.1. Word-entropy of languages

The average information content of words is represented by Entropy (Bentz and Alikaniotis, 2016). This metric would be higher for languages with a wider variety of word forms, i.e., languages that learn more details into word structure rather than a phrase or sentence structure.

Table 10
Adequacy and fluency scales.

Scales	Adequacy	Fluency
1	none	incomprehensible
2	little meaning	disfluent
3	much meaning	non-native
4	most meaning	good
5	all meaning	flawless

Table 11
Human evaluation summary.

Model			BHO→HI	MAG→HI	HI→BHO	HI→MAG
A	NS-1	AA	4.02	3.05	2.18	3.48
		AF	3.00	2.12	2.16	3.12
	NS-2	AA	3.98	3.05	2.32	3.78
		AF	2.78	2.98	2.04	3.56
	Agreement	K_A	0.725	0.950	0.730	0.687
		K_F	0.690	0.987	0.710	0.642
B	NS-1	AA	3.36	3.91	2.04	3.12
		AF	2.70	2.92	2.04	2.34
	NS-2	AA	3.39	4.02	1.96	4.24
		AF	2.64	3.02	1.92	2.16
	Agreement	K_A	0.812	0.672	0.725	0.540
		K_F	0.750	0.720	0.680	0.610
C	NS-1	AA	4.08	3.70	2.00	2.85
		AF	3.42	3.28	2.09	2.18
	NS-2	AA	3.96	3.64	2.22	2.34
		AF	3.06	3.19	2.26	2.08
	Agreement	K_A	0.752	0.800	0.512	0.462
		K_F	0.612	0.600	0.575	0.587
D	NS-1	AA	3.37	3.00	2.32	2.40
		AF	2.73	2.79	2.42	1.96
	NS-2	AA	3.40	3.34	2.14	2.56
		AF	2.63	2.61	2.14	2.10
	Agreement	K_A	0.800	0.745	0.637	0.437
		K_F	0.750	0.712	0.650	0.512

Note- AA: Average Adequacy, AF: Average Fluency, K_A: Inter-evaluator agreement Kappa coefficient for adequacy, K_F: Inter-evaluator agreement Kappa coefficient for fluency, NS-1: Native Speaker-1, NS-2: Native Speaker-2.

Let C be a text drawn from a vocabulary $Z = \{v_1, v_2, \ldots, v_k\}$ of size k. Furthermore, let word type probabilities are distributed according to $p(v) = P_r(v \in C)$ for $v \in Z$. The average information content of the word types is calculated by Shannon (1948) method as follows:

$$H(C) = -\sum_{j=1}^{k} p(v_j) \log_2(p(v_j)). \quad (9)$$

Entropy with higher values indicates language having high lexical richness as shown in Table 8. Translation direction for Hindi→Bhojpuri and Hindi→Magahi is from low to high lexical rich language. Thus, high lexical richness of target language compared to source language degrades the model's performance for Hindi→Bhojpuri and Hindi→Magahi compared to Bhojpuri →Hindi and Magahi→Hindi as shown in Table 6. According to Table 8, Magahi and Hindi have entropy scores of 5.1480 and 5.1474, respectively, indicating that Magahi is morphologically close to Hindi. Because this score is not statistically significant, we also compute the Type-to-Token Ratio (TTR) described in the next section, which reveals a substantial difference between Hindi and Magahi. We assess Type-to-Token Ratio at the word level to determine morphological complexity (Mundotiya et al., 2021).

5.5.2. Type-to-Token Ratio of languages

To calculate morphological complexity, we consider the ratio of word types over word tokens (Kettunen, 2014). The spectrum of word forms is expanded by using productive morphological markers. As a result, higher TTR value implies higher morphological complexity. Given a text C drawn from a vocabulary of word types $Z = \{v_1, v_2, \ldots, v_k\}$, the measure is written as follows:

$$TTR(C) = \frac{k}{\sum_{j=1}^{k} f(q_j)}, \quad (10)$$

Table 12
Examples of translated sentences

Languages	Sentences					
Bhojpuri (SRC)	e waraha se kenxra sarakAra xvArA anuxAna aura yojanA se AXAriwa sahAyawA se alaga kara-aMwaraNa ke xiSA meM baxalAva kalla galla ha.					
Hindi (TRG)	isa prakAra, kenxra sarakAra se anuxAnoM Ora yojanAoM se AXAriwa sahAyawA se hatakara kara-aMwaraNa kI xiSA meM baxalAva kiyA gayA hE.					
Language pairs	**Models**	**Translated Sentences**	A_1	F_1	A_2	F_2
Bhojpuri→Hindi	A	e. prakAra se keMxra sarakAra xvArA anuxAna aura yojanA se AXAriwa sahAyawA se alaga kara-aMwaraNa ke xiSA meM baxalAva kalla cale jAla hEM.	3	4	3	3
	B	e pUra se kena xa ra sarakAra xa vArA anuxAna aura yojanA se AXAriwa sahAyawA se alaga kara-aMwaraNa kI xiSA meM baxalAva yA kalla galla ha.	4	4	4	3
	C	e. waraha se keMxra sarakAra xvArA anuxAna aura yojanA se AXAriwa sahAyawA se alaga kara-aMwaraNa ke xiSA meM baxalAva kalla gayIla hE.	4	3	3	3
	D	e usakI se ken xa ra sarakAra xa vArA anuxAna aura yojanA se AXAriwa kI sahAyawA se alaga kara-aMwaraNa ka yA kI xiSA meM baxalAva kalla galla ha.	4	3	4	3
Languages	**Sentence**					
Magahi (SRC)	biratena-BArawa paraxyogikI BaglxArI hamanI ke saMjukwa xqsti Au hamanI ke mOjUxA Au BAbI pIDZI ke samqxXi ke mUla AXAra he.					
Hindi (TRG)	britena-BArawa prOxyogikI BAglxArI hamAre saMyukwa xqRti Ora hamArI mOjUxA waWA BAvI pIDZI kI samqxXi kA mUla AXAra hE.					
Language pairs	**Models**	**Translated Sentences**	A_1	F_1	A_2	F_2
Magahi→Hindi	A	biratena-BArawa paraxyogikI BaglxArI hamanI ke saMjuk wa xqsti Au hamanI kyA mOjUxA Au BAbI pIDZI kyA samqxXi kyA mUla AXAra hE.	4	3	3	3
	B	biratena-BArawa paraxa yogikI ke ka yA saMjuka BaglxArI hamanI qsa ka yA meM A hamanI BAbI A mOjUxA pIDa ka yA samqxa Xi kA mUla AXAra yaha.	3	3	4	2
	C	biratena-BArawa paraxyogikI BaglxArI hamanI ke saMjuk wa xqsti Ane kI hamanI yaha mOjUxA Ane vAlI BAbI pIDZI kI samqxXi ke mUla AXAra he.	3	3	3	3
	D	biratena-BArawa para xa yogikI BaglxArI hamanI ka yA saMjuka wa xqsa ti A hamanI ka yA mOjUxA A BAbI pIDa ka yA samqxa Xi ka yA mUla AXAra he.	4	3	3	2
Language	**Sentence**					
Hindi (SRC)	makAna mAlika re apanI hI musIbawoM meM PaMsA hE.					
Bhojpuri (TRG)	makAna mAlika re apanahI musIbawana meM Pazsala hava.					
Language pairs	**Models**	**Translated Sentences**	A_1	F_1	A_2	F_2
Hindi→Bhojpuri	A	Gara mAlika re APnA nE samasyAmA PazsAekA Can.	2	2	2	2
	B	makAna mAlika re apanI nE musIbawoM meM PaMsA Ca.	4	3	3	2
	C	makAna mAlika re APno nE musIbawamA PaMsA Ca.	2	2	2	2
	D	makAna mAlika re apanA nE musIbawoM meM PaMsA.	3	2	2	2
Language	**Sentence**					
Hindi (SRC)	mEM sawsaMga meM bETA WA, mEM sUpa kA hakaxAra hUz.					
Magahi (TRG)	hama sawasaMga meM baiTala halI,hama supa ke hakaxAra hI.					
Language pairs	**Models**	**Translated Sentences**	A_1	F_1	A_2	F_2
Hindi→Magahi	A	ma sawsafgamA baseko Wiez, ma supako hakaxAra Cu.	2	2	3	2
	B	ma sawa saMga basirahekA Wie, ma sUpa hakaxAra hUz.	2	2	2	2
	C	ma sawsafgamA basiraheko Wiez, ma supako hakaxAra Cu.	2	2	2	2
	D	hama pa safa meM bETA Wiyo, hama sUpa kA hakaxAra hUz.	3	2	2	2

Note-1: The scripts of Hindi, Bhojpuri and Magahi language are represented in WX-notation. *Note-2*: A_i: Adequacy score by i^{th} native speaker, F_i: Fluency score by i^{th} native speaker.

where, $f(q_j)$ is the token frequency of the j^{th} type.

According to Table 8, Bhojpuri and Magahi are more morphologically complex than Hindi, supporting the linguistic claim made for these languages. In Hindi→ Bhojpuri and Hindi→Magahi, the translation direction is from low to high significant complex language. The high complexity of target languages leads the model to be uncertain in learning the representation of vectors. Thus, the score of Hindi→Bhojpuri and Hindi→Magahi are less than Bhojpuri→Hindi and Magahi→Hindi.

5.6. Scalability test

To perform a scalability test, we randomly subsample the synthetic parallel training corpus of Bhojpuri→Hindi and Magahi→Hindi 5 times, discarding nearly half of the data at each step. Byte-pair embedding segmentation is learned on the total training corpus via sentencepiece. We set the frequency threshold for subword units to 10 in each subcorpus. Results of trained model for each subsample on Bhojpuri→Hindi and Magahi→Hindi are shown in Figs. 5 and 6. The consistent improvement on each subsample concludes that our model is robust with different data size. In Fig. 5, we see that variant *B* and *D* of TLSPG approach are showing near about the same performance up to 30000 data size (number of training sentences). After 30000 data size, variant *B* outperforms all the models for the Bhojpuri→Hindi language pair. The variants *A* and *C* go constant after 75000 sentences. In Fig. 6, variants *B* and *D* show constant improvement up to 135000 data size. After 135000 data size, variant *D* outperforms all the models for Magahi→Hindi language pair. The variants *A* and *C* show consistent improvement. Hence, variants *B* and *D* perform better on both the language pairs.

5.7. Using LSTM in-place of transformer in TLSPG

In addition to the above analysis, we have also conducted a study by replacing some components of the proposed approach. This replacement shows how the approach gets affected without the particular part. Here, we have replaced the Transformer part of the TLSPG approach with Long Short Term Memory *(LSTM)* architecture. We have performed experiments on four models: LSSL (replacing transformer in TSL), A_{lstm} (replacing transformer in variant *A*), C_{lstm} (replacing transformer in variant *C*) and D_{lstm} (replacing transformer in variant *D*). Results on all the four models are listed in Table 9. We have observed that replacing the transformer with LSTM hurts the performance of all the models. This study shows the effectiveness of the transformer in the proposed TLSPG approach.

5.8. Human evaluation

We have performed human evaluation on the translated sentence produced by all the four variants of TLSPG for Bhojpuri↔Hindi and Magahi↔Hindi language pairs. We have randomly selected 100 sentences from each language pair and evaluated adequacy and fluency scores by two native language speakers for each language pair based on the five-point scales as listed in Table 10. We have reported four types of scores- Average Adequacy (*AA*), Average Fluency (*AF*), Kappa coefficient for adequacy (K_A) and Kappa coefficient for fluency (K_F) on selected sentences in Table 11. Kappa coefficient helps in computing the inter-evaluator agreement between the judgement of different native speakers (Cohen, 1960). Based on the agreement computed for K_A and K_F between the two native speakers, we have observed that more than 60% of the judgements are similar for both Bhojpuri→Hindi and Magahi→Hindi language pairs. After going through *AA* and *AF* values, we have also seen that translated sentences contain sufficient semantic information of source sentences. Moreover, we have also included some examples of sentences on different variants of the TLSPG approach with their adequacy and fluency scores in Table 12 for better readability.

6. Conclusions

In this work, we have proposed TLSPG, a transfer learning-based semi-supervised pseudo-corpus generation approach for zero-shot translation systems. We have demonstrated the effectiveness of the proposed model on Bhojpuri ↔Hindi and Magahi↔Hindi language pairs in four different directions. The proposed approach outperformed the state-of-the-art models with +15.56 on Bhojpuri→Hindi, +8.13 on Magahi→Hindi, +3.98 on Hindi→Bhojpuri and +2 on Hindi→Magahi language pairs, respectively. We have also conducted various experiments using corpus-based approaches and human evaluations to support the performance of our model. Moreover, we have also evaluated the scalability test to show the robustness of the proposed model on different data sizes.

In the future, we will expand our approach to adversarial techniques for data augmentation and other zero-resource and extremely low-resource languages. We would also like to expand our approach to applications-based systems such as barrier-less communication between local speakers (e.g., Bhojpuri and Magahi) and medical practitioners in the healthcare domain. Furthermore, the developed model could also be adopted for mobile health applications exploiting low-resource environments in developing countries (Chib et al., 2015).

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

We would like to thank the Associate Editor and anonymous reviewers for insightful comments that helped us improve the quality of the manuscript significantly. The support and the resources provided by PARAM Shivay Facility under the National Supercomputing Mission, Government of India at the Indian Institute of Technology, Varanasi are gratefully acknowledged. The work of Ajay Pratap is partially supported by the Science and Engineering Research Board (SERB), Government of India under Grant SRG/2020/000318.

References

Abercrombie, G., 2016. A rule-based shallow-transfer machine translation system for Scots and English. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 578–584.

Al-Shedivat, M., Parikh, A., 2019. Consistency by agreement in zero-shot neural machine translation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1184–1197.

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., Macherey, W., 2019a. The missing ingredient in zero-shot neural machine translation. arXiv preprint arXiv:1903.07091.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M.X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., Wu, Y., 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S.,

Post, Zampieri, M., 2019. Findings of the 2019 conference on machine translation (WMT19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Association for Computational Linguistics, Florence, Italy, pp. 1–61. https://doi.org/10.18653/v1/W19-5301.

Bentz, C., Alikaniotis, D., 2016. The word entropy of natural languages. arXiv:1606.06996.

Calixto, I., Stein, D., Matusov, E., Castilho, S., Way, A., 2017. Human Evaluation of Multi-modal Neural Machine Translation: A Case-Study on E-Commerce Listing Titles, in: Proceedings of the Sixth Workshop on Vision and Language, Association for Computational Linguistics, Valencia, Spain. pp. 31–37. https://aclanthology.org/W17-2004, 10.18653/v1/W17-2004.

Chib, A., van Velthoven, M.H., Car, J., 2015. mHealth Adoption in Low-Resource Environments: A Review of the Use of Mobile Healthcare in Developing Countries. Journal of Health Communication 20, 4–34. https://doi.org/10.1080/10810730.2013.864735.

Chu, C., Wang, R., 2018. A survey of domain adaptation for neural machine translation. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1304–1319.

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20 (1), 37–46. https://doi.org/10.1177/001316446002000104.

Cooper Stickland, A., Berard, A., Nikoulina, V., 2021. Multilingual Domain Adaptation for NMT: Decoupling Language and Domain Information with Adapters. In: Proceedings of the Sixth Conference on Machine Translation. Association for Computational Linguistics, Online, pp. 578–598.

Dabre, R., Chu, C., Kunchukuttan, A., 2020. A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR) 53, 1–38.

Dabre, R., Fujita, A., Chu, C., 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. 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). Association for Computational Linguistics, Hong Kong, China, pp. 1410–1416. https://doi.org/10.18653/v1/D19-1146.

Diwakar, S., Goyal, P., Gupta, R., 2010. Transliteration among indian languages using wx notation. In: Proceedings of the Conference on Natural Language Processing 2010. Saarland University Press, pp. 147–150.

Edunov, S., Ott, M., Auli, M., Grangier, D., 2018. Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp. 489–500. https://doi.org/10.18653/v1/D18-1045.

Fei, H., Zhang, Y., Ren, Y., Ji, D., 2021. Optimizing attention for sequence modeling via reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 1–10. https://doi.org/10.1109/TNNLS.2021.3053633.

Firat, O., Sankaran, B., Al-onaizan, Y., Yarman Vural, F.T., Cho, K., 2016. Zero-resource translation with multi-lingual neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas. pp. 268–277. https://aclanthology.org/D16-1026, 10.18653/v1/D16-1026.

Ghaddar, A., Langlais, P., 2020. SEDAR: a Large Scale French-English Financial Domain Parallel Corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association Marseille, France, pp. 3595–3602.

Gu, J., Wang, Y., Cho, K., Li, V.O., 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1258–1268.

Heafield, K., 2011. KenLM: Faster and smaller language model queries, in: Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Edinburgh, Scotland. pp. 187–197. https://www.aclweb.org/anthology/W11-2123.

Hokamp, C., Glover, J., Gholipour Ghalandari, D., 2019. Evaluating the supervised and zero-shot performance of multi-lingual translation models, in: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 209–217.

Hurskainen, A., Tiedemann, J., 2017. Rule-based machine translation from English to Finnish, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark. pp. 323–329. https://aclanthology.org/W17-4731, 10.18653/v1/W17-4731.

Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J., 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for. Computational Linguistics 5, 339–351. https://doi.org/10.1162/tacl_a_00065. https://aclanthology.org/Q17-1024.

Kettunen, K., 2014. Can type-token ratio be used to show morphological complexity of languages? Journal of Quantitative Linguistics 21, 223–245.

Klavans, J., Morgan, J., LaRocca, S., Micher, J., Voss, C., 2018. Challenges in Speech Recognition and Translation of High-Value Low-Density Polysynthetic Languages. In: Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track), Association for Machine Translation in the Americas Boston, MA, pp. 283–293.

Koehn, P., 2009. Statistical machine translation. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic, pp. 177–180.

Kumar, A., Mundotiya, R.K., Singh, A.K., 2020. Unsupervised approach for zero-shot experiments: Bhojpuri-Hindi and Magahi–Hindi@LoResMT 2020, in: Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, Association for Computational Linguistics, Suzhou, China. pp. 43–46. https://aclanthology.org/2020.loresmt-1.6.

Lakew, S.M., Federico, M., Negri, M., Turchi, M., 2018. Multilingual neural machine translation for low-resource languages. IJCoL. Italian Journal of Computational Linguistics 4, 11–25.

Lakew, S.M., Negri, M., Turchi, M., 2021. Self-learning for zero shot neural machine translation. arXiv preprint arXiv:2103.05951.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. 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. Association for Computational Linguistics, Onlin, pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L., 2020. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics 8, 726–742. https://doi.org/10.1162/tacl_a_00343.

Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., Sun, J., 2018. A neural interlingua for multilingual machine translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 84–92.

Mi, C., Xie, L., Zhang, Y., 2020. Improving adversarial neural machine translation for morphologically rich language. IEEE Transactions on Emerging Topics in Computational Intelligence 4, 417–426. https://doi.org/10.1109/TETCI.2019.2960546.

Mundotiya, R.K., Singh, M.K., Kapur, R., Mishra, S., Singh, A.K., 2021. Linguistic resources for bhojpuri, magahi, and maithili: Statistics about them, their similarity estimates, and baselines for three applications. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20. https://doi.org/10.1145/3458250, DOI: 10.1145/3458250.

Mutal, J., Bouillon, P., Gerlach, J., Estrella, P., Spechbach, H., 2019. Monolingual backtranslation in a medical speech translation system for diagnostic interviews - a NMT approach, in: Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks, European Association for Machine Translation, Dublin, Ireland. pp. 196–203.

Nasution, A.H., Syafitri, N., Setiawan, P.R., Suryani, D., 2017. Pivot-based hybrid machine translation to support multilingual communication. In: 2017 International Conference on Culture and Computing (Culture and Computing), pp. 147–148. https://doi.org/10.1109/Culture.and.Computing.2017.22.

Och, F.J., Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational Linguistics 29, 19–51. https://doi.org/10.1162/089120103321337421. https://www.aclweb.org/anthology/J03-1002.

Ojha, A.K., Malykh, V., Karakanta, A., Liu, C.H., 2020. Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In: Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages. Association for Computational Linguistics, Suzhou, China, pp. 33–37.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M., 2019. 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 (Demonstrations). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 48–53. https://doi.org/10.18653/v1/N19-4009.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp. 311–318.

Pham, N.Q., Niehues, J., Ha, T.L., Waibel, A., 2019. Improving zero-shot translation with language-independent constraints. In: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pp. 13–23.

Popović, M., 2015. chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal. pp. 392–395. 10.18653/v1/W15-3049.

Post, M., 2018. A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium. pp. 186–191. 10.18653/v1/W18-6319.

Sennrich, R., Haddow, B., Birch, A., 2016. Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany. pp. 1715–1725. https://aclanthology.org/P16-1162, 10.18653/v1/P16-1162.

Sestorain, L., Ciaramita, M., Buck, C., Hofmann, T., 2018. Zero-shot dual machine translation. arXiv preprint arXiv:1805.10338.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Skianis, K., Briand, Y., Desgrippes, F., 2020. Evaluation of Machine Translation Methods applied to Medical Terminologies, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis,

Association for Computational Linguistics, Online. pp. 59–69. https://aclanthology.org/2020.louhi-1.7, 10.18653/v1/2020.louhi-1.7.

Snover, M.G., Dorr, B.J., Schwartz, R.M., Micciulla, L., Makhoul, J., 2006. 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, AMTA 2006, Cambridge, Massachusetts, USA, August 8–12, 2006, Association for Machine Translation in the Americas. pp. 223–231. https://aclanthology.org/2006.amta-papers.25/.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215.

Tiedemann, J., 2012. Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA) Istanbul, Turkey, pp. 2214–2218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Zhang, B., Williams, P., Titov, I., Sennrich, R., 2020. Improving massively multilingual neural machine translation and zero-shot translation. arXiv preprint arXiv:2004.11867.