Natural Language Engineering (2021), **27**, pp. 271–292
doi:10.1017/S1351324920000303

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Neural machine translation of low-resource languages using SMT phrase pair injection

Sukanta Sen[1,*], Mohammed Hasanuzzaman[2], Asif Ekbal[1], Pushpak Bhattacharyya[1] and Andy Way[2]

[1]Indian Institute of Technology Patna, India and [2]ADAPT Centre, Dublin City University, Ireland
*Corresponding author. E-mail: sukanta.pcs15@iitp.ac.in

(Received 12 February 2019; revised 22 April 2020; accepted 24 April 2020; first published online 17 June 2020)

Abstract

Neural machine translation (NMT) has recently shown promising results on publicly available benchmark datasets and is being rapidly adopted in various production systems. However, it requires high-quality large-scale parallel corpus, and it is not always possible to have sufficiently large corpus as it requires time, money, and professionals. Hence, many existing large-scale parallel corpus are limited to the specific languages and domains. In this paper, we propose an effective approach to improve an NMT system in low-resource scenario without using any additional data. Our approach aims at augmenting the original training data by means of parallel phrases extracted from the original training data itself using a statistical machine translation (SMT) system. Our proposed approach is based on the gated recurrent unit (GRU) and transformer networks. We choose the Hindi–English, Hindi–Bengali datasets for Health, Tourism, and Judicial (only for Hindi–English) domains. We train our NMT models for 10 translation directions, each using only 5–23k parallel sentences. Experiments show the improvements in the range of 1.38–15.36 BiLingual Evaluation Understudy points over the baseline systems. Experiments show that transformer models perform better than GRU models in low-resource scenarios. In addition to that, we also find that our proposed method outperforms SMT—which is known to work better than the neural models in low-resource scenarios—for some translation directions. In order to further show the effectiveness of our proposed model, we also employ our approach to another interesting NMT task, for example, old-to-modern English translation, using a tiny parallel corpus of only 2.7K sentences. For this task, we use publicly available old-modern English text which is approximately 1000 years old. Evaluation for this task shows significant improvement over the baseline NMT.

Keywords: Machine translation; Translation technology

1. Introduction

Neural machine translation (NMT) (Forcada and Ñeco 1997; Kalchbrenner and Blunsom 2013; Cho *et al*. 2014; Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015) has recently drawn significant attention to the researchers due to its encouraging performance on publicly available benchmark datasets (Bojar *et al*. 2016) and rapid adoption in the production systems (Wu *et al*. 2016; Crego *et al*. 2016; Junczys-Dowmunt, Dwojak, and Hoang 2016). The key points of NMT are it generates fluent outputs and it can be implemented as a single end-to-end neural system unlike long-dominant phrase-based Statistical Machine Translation (SMT) (Koehn, Och, and Marcu 2003) which combines many submodules. The performance of an NMT system largely depends on the amount of parallel data we have. It produces good translations when we have sufficient training data; however, it performs poorly when the training data size is insufficient. The size of this sufficient data for NMT training is in the order of millions of parallel sentences

(Lample *et al.* 2018). In contrast to NMT, SMT models are inherently known to be better than NMT in the absence of enough training data.

Although SMT performs better than NMT in the absence of large parallel corpus, there has been a growing interest among the researchers to build effective NMT models in such scenarios as well. One reason that makes NMT a better choice, even in the absence of sufficient data, is that NMT makes a huge jump in BiLingual Evaluation Understudy (BLEU) score as the data size increases, whereas SMT improves with a fixed rate (Koehn and Knowles 2017). NMT requires a huge amount of parallel data for building a good translation system, and absence of such corpora makes NMT suffer from the adequacy problem (Koehn and Knowles 2017).

The quality of an NMT system heavily depends on the training data size. The standard systems make use of parallel corpus having millions of sentences. However, it is not that we only lack of training data for certain language pairs, but many a times we also face with the problem of low-resource scenario for many domains such as medical, tourism, and judicial. In that case, translation again becomes a challenging task because of the absence of the parallel data for those domains. For example, many Indic languages do not have enough parallel corpus required to build the robust NMT systems. Only few thousands of parallel sentences are available (Jha 2010). In the absence of sufficient amount of data, a model learns poorly because of the low counts of source–target units. One of the major challenges of NMT, irrespective of the training data size, is handling of the rare words. However, if the data size is very small then most of the source–target pairs occur in very less number.

In this work, we propose a method to substantially improve the NMT system for low-resource languages and/or domains. We extract phrase pairs from the original training data using a phrase-based SMT (Koehn *et al.* 2003) training and augment the original training corpus by adding the most probable pairs as parallel sentence pairs. By phrase, we do not necessarily mean any linguistic phrase—it is rather a consecutive sequence of words. We evaluate our approach using the BLEU score (Papineni *et al.* 2002) against the baseline models constructed using the standard attention-based (Bahdanau *et al.* 2015) gated recurrent unit (GRU) (Cho *et al.* 2014) and transformer-based (Vaswani *et al.* 2017) NMT systems. Our experiments show that our proposed model attains significant performance gains over the baseline models under a very low-resource scenario. Our approach is different from the existing approaches in the following ways: (i) our system makes use of a relatively smaller corpus consisting of only 5–23k parallel sentences and (ii) we include the phrase pairs directly in the training corpus treating as sentence pairs.

We summarizes the key contributions and/or characteristics of our proposed approach as follows.

- We propose an effective NMT model with feedback from SMT phrases for translating low-resource languages.
- We empirically establish that our proposed approach improves the performance of NMT system under low-resource scenario, showing improvements over the baselines for the English–Hindi and Hindi–Bengali language pairs.
- We empirically show that transformer works significantly better than the attention-based GRU in low-resource scenarios.
- We also build an NMT system for old-to-modern English translation using our proposed approach and observe significant improvement over the baseline. Main intuition of doing this was to establish how generic and effective our proposed approach is for translating the texts of completely different genres and structures.

The remainder of the paper is organized as follows. In Section 2, we define the problem and present the underlying motivation of our current work. Section 3 presents an overview of the existing literature. In Section 4, we describe the proposed method. Sections 5 and 6 discuss the datasets and experimental setup, respectively. In Section 7, we report the results along with proper analysis. Finally, in Section 8, we conclude with future work directions.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

2. Problem definition and motivation

Both NMT and SMT require a large-scale high-quality parallel corpus for training a good-quality machine translation (MT) system. Absence of such corpora makes NMT suffer from the adequacy problem (Koehn and Knowles 2017). Creating a high-quality large-scale parallel corpus is expensive as it requires time, money, and professionals to translate a large amount of texts. As a result, many of the existing large-scale parallel corpora are limited to some specific languages and domains.

The quality of any MT system can be characterized by its adequacy and fluency. The long-dominant SMT has been found to be good at handling adequacy, but lacks in fluency. Recently, NMT has become the new state of the paradigm to MT. However, it has been reported that NMT sacrifices adequacy at the cost of fluency (Koehn and Knowles 2017). The performance of an NMT system greatly depends on the amount of parallel corpus: more the data better is the performance. However, having sufficient corpus for training an NMT system is a challenge. The adequacy is a direct measure of how well an NMT system learns mapping between the symbols in source and target languages. However, NMT fails to capture these mappings when the parallel corpus is not enough. This is often the case that sufficient parallel corpus is not available for many language pairs as well as for some restricted domains. So, in order to help the NMT models learn the mapping between words in the absence of sufficient parallel corpus, we extract phrases from the original training data and add to it (i.e., the original training corpus) for better evidences. This, in turn, provides an implicit knowledge about the mappings between the source–target pairs.

In our current work, we propose an effective approach for the translation of a variety of texts and languages. Phrases extracted from SMT are fed as input to the training of NMT. Firstly, we build the NMT systems for the resource-scarce Indian languages, and secondly, we translate the old English texts to the modern. India is a multilingual country with great linguistic and cultural diversities. The resources and tools in the form of parallel corpus, morphological analyzers, parts-of-speech tagger, etc. are readily not available in the required measures. Translating old text to the modern text is very important for various purposes. Human languages are constantly evolving and changing over time to reflect sociocultural changes, fit current conventions, mores, expressions, and needs. This change in a language often requires "rewriting" the old texts for the modern readers in the same language. In line with global trends, old texts are increasingly available in the forms that computer can process. These ever expanding records (e.g., historical records, scanned books, academic papers, large-scale corpora, and maps)—either digitally born or reconstructed through digitization pipelines—are too big to be "rewritten" manually. We pose this rewriting of old text as an MT problem and use our proposed method to improve this rewriting.

3. Related work

Having enough parallel corpus is a big challenge in NMT, and it is very unlikely to have millions of parallel sentences for every language pair. A few attempts have been made to build NMT systems for the low-resource language pairs (Sennrich, Haddow, and Birch 2016a; Zhang and Zong 2016; Gulcehre *et al.* 2017), which incorporated huge monolingual corpus in the source and target sides.

Sennrich *et al.* (2016a) have incorporated monolingual data on the target side to investigate two methods of filling the source side of the monolingual data. In the first method, they have used a dummy source sentence for every target sentence, and in the second method, they used a synthetic source sentence obtained via back translation. They claimed that the second method is more effective. However, if there is not enough parallel data, quality of back translation is again a problem.

Zhang and Zong (2016) explored the effect of incorporating large-scale source-side monolingual data in NMT in different ways. In the first approach, inspired by Sennrich *et al.* (2016a), they first built a baseline model and then obtained parallel synthetic data by translating the monolingual data. These parallel data along with the original data are used for training an attention-based

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

GRU system. The second method used the multitask learning framework to generate the target translation and reorder the source-side sentences at the same time. They claimed that usage of source-side monolingual data in NMT is more effective than that of SMT.

Gulcehre *et al.* (2017) have proposed two alternative methods to integrate monolingual data on the target side, namely shallow fusion and deep fusion. In shallow fusion, the top K hypotheses (produced by NMT) in each time step t are re-scored using the weighted sum of the scores given by the NMT (trained on parallel data) and a recurrent neural network-based language model (RNNLM). Whereas in deep fusion, hidden states obtained at each time step t of RNNLM and NMT are concatenated and the output is generated from that concatenated state.

In recent times, Arthur, Neubig, and Nakamura (2016) have proposed a model to incorporate translation lexicons through calculating lexical predictive probability and adding this probability to the input of the Softmax. Feng *et al.* (2017) proposed a method that extracted phrase translation dictionary from the corpus using word alignment, and the phrase translation probability is used in the NMT model to construct the local memory.

Zoph *et al.* (2016) applied transfer learning for low-resource NMT. They trained a model on high-resource language pair, and then the learned parameters were used for training a low-resource language pair. However, it requires selecting both the high- and low-resource language pairs to be of similar types (i.e., closer to each other). So this approach may not work if the language pairs are distant. There are two fundamental differences between our proposed approach and theirs. Unlike theirs, we do not use any large amount of additional parallel corpus, rather we use a relatively smaller corpus.

Wang *et al.* (2018) proposed simple data augmentation technique through randomly replacing words in both the source sentence and the target sentence with other random words from their corresponding vocabularies. He *et al.* (2016) have incorporated SMT features such as translation model, language model (LM) under log-linear framework during the beam search of decoding step.

Wang *et al.* (2017), Wang, Tu, and Zhang (2018) have proposed NMT model advised by SMT, where at each decoding step, SMT offers additional recommendations and the recommendations are scored with a classifier for combining with the NMT model in an end-to-end manner. Zhao *et al.* (2018) also have used phrase table as recommendation through adding bonus to words worthy of recommendation, for helping NMT in predicting adequate words.

NMT always shows weakness in translating the rare words. Fadaee, Bisazza, and Monz (2017) have proposed an approach for handling rare words through data augmentation for English–German language pair. Their approach also made use of huge monolingual corpus for generating sentence pairs containing rare words, and these generated sentence pairs are used for training the NMT models. Although the said pair does not fall under the low-resource category, they created simulated low-resource settings to perform the experiments and claimed to have achieved substantial improvement on the BLEU score. However, in our experimental settings, we truly use the low-resource languages, for which having a large monolingual corpus is also a challenge.

Song *et al.* (2019) have investigated a data augmentation method for constraining NMT with pre-specified translations. In this method, source sentences in the training data are code-switched by replacing source phrases with their target translations allowing the model to learn lexicon translations by copying source-side target words. However, in our approach, we do not code-switch the training data, instead we use phrase pairs as training data.

Most of the earlier works related to NMT in low-resource scenario tried to incorporate monolingual data either in the source or target side. The effect of adding monolingual data in NMT is similar to that of building LM on a large-scale monolingual data in SMT. It makes output more fluent; however, NMT always lacks in generating adequate output. Adding monolingual data do not contribute much in improving the adequacy. A number of attempts related to low-resource NMT also tried to expand the training data by adding the back-translated monolingual data. At first, a model is trained using the available training data, and then the monolingual corpus is passed for

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

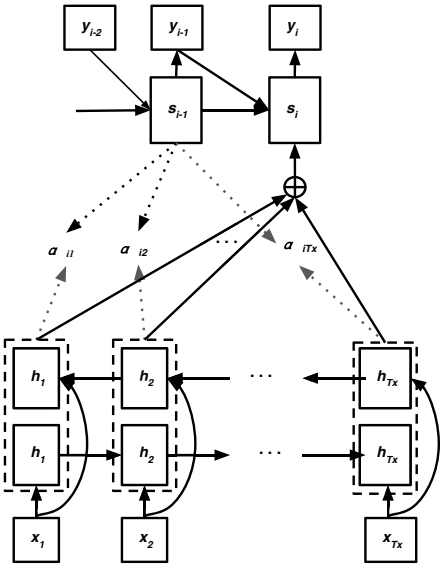

Figure 1. Attention-based GRU NMT architecture.

translation. Quality of the parallel data obtained from the translation of monolingual data depends on the size of the original parallel data. If the data size is very small, then the translated data may not help much. However, the effect of adding source–target phrases into the training data is less explored (Sen *et al.* in press).

Recently, unsupervised NMT (Lample *et al.* 2018; Artetxe, Labaka, and Agirre 2018; Ren *et al.* 2019; Lample and Conneau 2019), semi-supervised NMT (Zhang *et al.* 2018), and unsupervised pre-training (Ramachandran, Liu, and Le 2017) have been emerged and shown promising results on the related languages. These techniques require a huge amount of monolingual data. It has been shown that pure unsupervised technique does not work for distance language pairs, for example, the pairs we are dealing with in this work (Guzmán *et al.* 2019).

4. Proposed method

Our focus is on low-resource scenario, and in order to handle this situation, we add the phrase pairs extracted from the training corpus as feedback to the NMT framework during training. Our proposed approach is not specific to any NMT architecture. We perform experiments on two state-of-the-art neural networks, namely attention-based (Bahdanau *et al.* 2015) GRU and transformer (Vaswani *et al.* 2017) models. Here, we briefly describe the two networks and then present the details of the proposed method.

4.1 Attention-based GRU

The goal of NMT is to translate a sequence of source words into a sequence of target words with the help of a large neural network. The basic architecture of an NMT, shown in Figure 1, uses two recurrent neural networks, one is called encoder and other is known as the decoder. The encoder converts the source sentence into a dense fixed-length vector, and then the decoder generates target sentence from that vector. But the main drawback of this encoder–decoder approach is that it fails drastically as length of the input sentence grows. The encoder–decoder approach assumes that the encoder can encode the whole sentence into a fixed-length vector, which is not realistic,

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

specifically for the longer sentences. To mitigate this drawback, Bahdanau *et al.* (2015) came up with an idea which focused on the whole input sentence, while generating the outputs.

Formally, given a sequence of source words x $(= x_1, x_2, x_3, ..., x_{T_x})$ and the previously translated $i-1$ words y $(= y_1, y_2, y_3, ..., y_{i-1})$, the conditional probability of the ith output y_i is calculated as

$$p(y_i|\boldsymbol{x}, y_{<(i-1)}) = \text{softmax}(W_o t_i) \tag{1}$$

where t_i, the input to the softmax, is computed as

$$t_i = \tanh(W_s s_i + W_e y_{i-1} + W_c c_i) \tag{2}$$

where W_s, W_e, W_c, and W_o are the model parameters. The hidden state s_i in the decoder at time step i is computed as

$$s_i = g(s_{i-1}, y_{i-1}, c_i) \tag{3}$$

Here, g is a nonlinear transform function, which is usually a long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) or a GRU (Cho *et al.* 2014), and c_i is the context vector at time step i, which is calculated as a weighted sum of the input annotations h_j:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} \mathrm{h}_j \tag{4}$$

where T_x is the length of the source sequence and h_j is the encoder hidden state at jth time step and computed using a nonlinear transformation (such as GRU and LSTM) function as

$$h_j = f(h_{j-1}, x_j) \tag{5}$$

The normalized weight α_{ij} for h_j is calculated as

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{6}$$

$$e_{ij} = V_a^T \tanh(U_a s_{i-1} + W_a h_j) \tag{7}$$

where V_a, U_a, and W_a are the trainable parameters. All of the parameters in the NMT model are optimized to maximize the following conditional log-likelihood of the N parallel sentences

$$\ell(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{T_y} \log p(y_i|s_i, y_{i-1}, c_i) \tag{8}$$

where T_y is the length of the target sequence.

4.2 Transformer network

In recurrent network, the representation at time step i is dependent of previous time stamps. Vaswani *et al.* (2017) proposed the transformer network, block diagrammatic representation is shown in Figure 2, which completely depends on *self-attention* and removes the recurrent operations found in the previous NMT approach—allowing for parallelization of computing at all time stamps in encoder–decoder. However, in the absence of recurrence, to capture the token position within the input sentence, a positional encoding is added with each input embedding before passing to the encoder. The encoder consists of several identical layers. Each layer is composed of mainly two sublayers: a multihead self-attention layer and a position-wise feed-forward network layer. The decoder consists of several identical layers like encoder and operates in a similar to

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

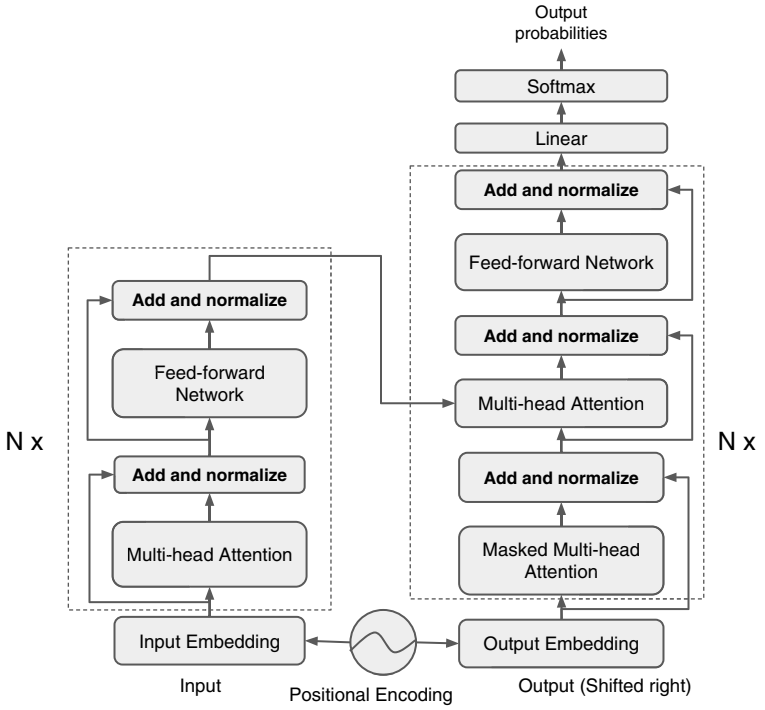

Figure 2. Transformer architecture.

encoder. In addition to the two sublayers in encoder, decoder inserts a third sublayer, which performs multihead attention over the output of the encoder. Each of these sublayers is followed by a layer called normalization. The decoder works similar to the decoder in Bahdanau *et al.* (2015) and generates one token at a time using a softmax layer. For more specific details of the network, please refer to Vaswani *et al.* (2017).

Mathematically, as in Vaswani *et al.* (2017), positional encoding is defined as

$$PE_{(pos,2i)} = \sin(pos/10,000^{2i/d}) \tag{9}$$

$$PE_{(pos,2i+1)} = \cos(pos/10,000^{2i/d}) \tag{10}$$

where pos is the position, and i is the ith dimension of a d dimensional input vector. Self-attention is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where Q, K, and V are the queries, keys, values packed together into matrices. d_k is the ratio of d to number of heads denoted as h. One multihead is a concatenation of multiple h heads, defined as

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{12}$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{13}$$

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

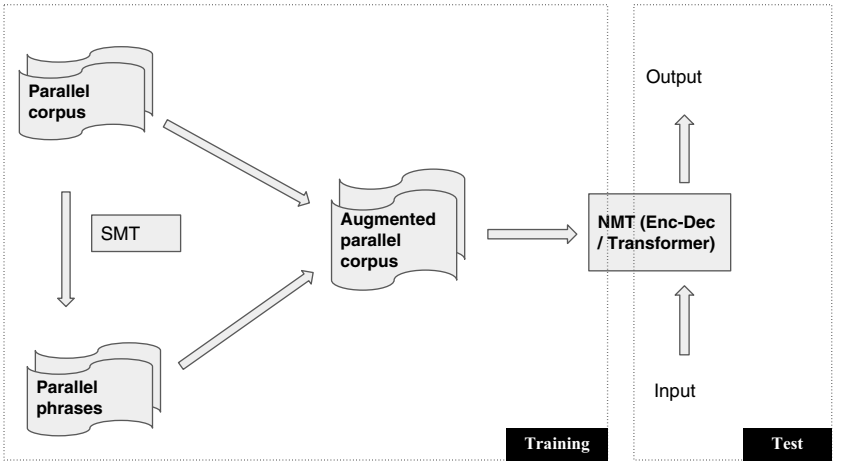

Figure 3. Our proposed phrase injection NMT approach.

where the projections are parameter matrices $W_i^Q, W_i^K, W^O \in \mathbb{R}^{d \times d_k}$. Feed-forward network in each layer of the encoder and decoder is formulated as

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{14}$$

where the projections are parameter matrices $W_1, W_2 \in \mathbb{R}^{d \times d}$.

4.3 Data augmentation

The NMT models are extremely data hungry, and in the absence of large training corpus, it does not learn the model parameters properly. In our work, we propose an approach for training NMT models using small corpora, especially under a situation for translating domain-specific small corpora.

The overall process flow of our architecture is depicted in Figure 3. The core idea is to provide more information about the alignment between the source and target phrases. When a sentence pair is passed through an encoder–decoder, the sentence pair does not carry any information about the mapping between the source and target phrases. The model learns the translation mappings implicitly by predicting and rectifying the error over a large parallel corpus. However, the model fails to learn the association between the phrases when the corpus size is small. So, apart from feeding sentence pairs into the network, we also feed phrase pairs as the training examples. This gives an illusion of having a larger corpus.

In order to perform this feedback mechanism, we first extract parallel phrases from the corpus and then add these parallel phrases in the training set. To extract parallel phrases, we use the Moses (Koehn *et al.* 2007) SMT system. We train a source–target phrase-based SMT (Koehn *et al.* 2003) and extract all the phrase pairs from the phrase table. Out of these, many parallel phrases are not sound, that is, there can be many incorrect source–target alignments (Koehn *et al.* 2003). We set different conditions while choosing the phrases from the phrase table. Assume that every source phrase e is aligned to a set of target phrases $F = (f_1, f_2, ..., f_n)$. Note that n may vary for each source phrase. So, for each source phrase e in the phrase table, we extract three sets of parallel phrases:

1. First set ($Set_{p \geq 0.5}$): set of parallel phrases (e, f_t) provided $P(f_t|e) \geq 0.5$;
2. Second set ($Set_{p=1.0}$): set of parallel phrases (e, f_t) provided $P(f_t|e) = 1.0$;
3. Third set (Set_{all}): for this set, we consider all the phrase pairs from the phrase table.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 1. Dataset statistics showing the number of sentences and tokens. By old, we refer to old English and by mod, we refer to modern English

		#Sent	#Token		#Sent	#Token		#Sent	#Token	
			English	Hindi		Hindi	Bengali		Old	Mod
Health	Train	23,000	395,859	413,124	22,012	392,993	318,967	–	–	–
	Test	1000	17,064	17,7626	956	16,902	13,709	–	–	–
	Dev	1000	17,006	17,746	950	16,764	13,600	–	–	–
Tourism	Train	23,000	392,557	390,463	21,950	371,415	302,726	–	–	–
	Test	1000	17,010	16,947	1000	16,929	13,870	–	–	–
	Dev	1000	16,731	16,693	1000	17,169	14,001	–	–	–
Judicial	Train	5561	134,745	144,244	–	–	–	–	–	–
	Test	1000	24,322	26,165	–	–	–	–	–	–
	Dev	1000	23,819	25,420	–	–	–	–	–	–
–	Train	–	–	–	–	–	–	2674	72,207	77,914
	Test	–	–	–	–	–	–	500	14,001	15,176
	Dev	–	–	–	–	–	–	500	14,349	15,638

Since the number of phrase pairs is larger than the number of original parallel sentences, to maintain a fair ratio between them, we use the following formula for combining phrase pairs with the original training set.

$$\text{Augmented corpus} = N \times \text{original corpus} + \text{extracted phrase pairs} \tag{15}$$

We combine the extracted set (of parallel phrases) with N times of the original corpus, where N is calculated as

$$N = \frac{\text{Number of extracted phrase pairs}}{\text{Number of original parallel sentences}}$$

Without this, training set will contain mostly the phrases, and as the phrases are smaller in length, they may make the model biased towards the phrase length.

5. Datasets

For experiments, we use English–Hindi and Hind–Bengali parallel corpora from the multilingual Indian Language Corpora Initiative (ILCI) (Jha 2010). The ILCI parallel corpora are from the two domains: Health (ILCI-H) and Tourism (ILCI-T), each of these comprising of 25k parallel sentences. These corpora have insufficient number of parallel sentences compared to the other language pairs found in literature. Indian languages do not have sufficient corpus required to train an NMT system, and thus they fall under the low-resource category as the standard NMT requires millions of parallel sentences for training. For experimentation, we randomly split each corpus into three sets: Train, Test, and Dev. Details are shown in Table 1.

For judicial domain data, we use IIT Bombay English-Hindi parallel corpus (Kunchukuttan, Mehta, and Bhattacharyya 2018). It consists of parallel sentences from miscellaneous domains and out of which only 7561 parallel sentences belong to the judicial domain. For experiments, we

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 2. Vocabulary size for different language pairs

	English↔Hindi		Hindi↔Bengali		O→M English	
Health	17,141	21,434	20,903	24,623	–	–
Tourism	24,861	28,168	27,411	33,080	–	–
Judicial	9795	8777	–	–	–	–
–	–	–	–	–	8878	5102

O, Old; M, Modern.

Table 3. Training data sizes for different models after adding phrases

	Health		Judicial	Tourism		O→M
Model	E ↔ H	H ↔ B	E ↔ H	E ↔ H	H ↔ B	E
Baseline	23,000	22,012	5561	23,000	21,950	2674
+ $Set_{p\geq 0.5}$	635,085	1,442,262	239,227	642,446	1,475,416	385,015
+ $Set_{p=1.0}$	457,103	1,218,393	165,498	486,240	1,258,791	341,659
+ Set_{all}	1,516,959	1,973,514	511,360	1,641,803	1,926,374	485,739

E, English; H, Hindi; B, Bengali; O, old; M, modern.

randomly split these judicial domain parallel sentences into Train, Dev, and Test sets consisting of 5561, 1000, and 1000, respectively.

As old English texts, we use the publicly available *The Homilies of the Anglo-Saxon Church*[a] by Ælfric of Eynsham (c.950–c.1010) who was a prolific author in old English and its translation by Benjamin Thorpe (c.1782–c.1870) as modern English texts. We call it Old English-Modern English (OE-ME) corpus.

The OE–ME corpus is tiny in size, and it has 720 parallel paragraphs in 40 sections. Most of the parallel paragraphs have equal number of OE–ME parallel sentences which help in aligning the parallel sentences. Some parallel paragraphs are discarded as they do not have equal number of OE–ME sentences to avoid misalignment which gives rise to a total of 3716 parallel sentences. We do not use any sentence aligner as only a few sentences are discarded. We randomly split it into three sets: Train, Test, and Dev containing 2716, 500, and 500 sentences, respectively. For tokenization, we use *tokenizer.perl* which is part of the Moses SMT system. Details of the datasets are presented in Table 1.

6. Experimental setup

Attention-based GRU models: We use Nematus (Sennrich *et al.* 2017) for training the NMT models. Our neural models are trained on word level. We create vocabulary from the training set for the different systems. The size of the vocabularies used in training the models is shown in Table 2. The augmented data size for each model are shown in Table 3. We set the embedding size as 128, hidden size as 256, and the learning rate as 0.001. Note that we tried higher embedding and hidden dimensions but did not work as the training data size is very small. Encoder and decoder are two-layered GRU blocks. The models are trained with mini-batch size of 40, and we restrict the maximum sentence length to 80. We use the Adam optimizer (Kingma and Ba 2015) for optimizing the models. The training stops on meeting the early stopping criteria. We use the

[a]https://en.wikisource.org/wiki/The_Homilies_of_the_Anglo-Saxon_Church.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

early stopping based on BLEU measure with an early stopping patience value 10. All the models run for 110–130k (approx.) updates before the early stopping. For decoding, we set the beam size as 3. For the other parameters, default values were used.

Transformer-based models: For training the models, we use Sockeye (Hieber *et al.* 2017), a toolkit for NMT. We set default the embedding dimension of 512, hidden dimension of 512, learning rate of 0.0002, and dropout rate of 0.2. Number of layers in each of encoder and decoder is 6. Number of multihead attention is 8. We use Adam optimizer (Kingma and Ba 2015) optimizer. We keep a mini-batch size of 2000 words.[b]

Phrase extraction: We use the Moses (Koehn *et al.* 2007) toolkit for training a phrase-based SMT system. The phrase table, generated during training, is used for extracting the phrases. For training, we keep the following settings in the Moses: grow-diag-final-and heuristics for word alignment, msd-bidirectional-fe for reordering model, and 4-gram LM with modified Kneser–Ney smoothing (Kneser and Ney 1995) using KenLM (Heafield 2011). However, we note that the order of LM does not affect the phrase table.

We train the following three types of NMT models for each of health, tourism, and judicial domain corpora.

1. *Baseline*: The NMT model is trained only on the original parallel corpus.
2. *Baseline* + $Set_{p\geq 0.5}$: The NMT is trained on the original parallel corpus along with phrase pair set $Set_{p\geq 0.5}$ (see Section 4.3).
3. *Baseline* + $Set_{p=1.0}$: The NMT is trained on the original parallel corpus along with phrase pair set $Set_{p=1.0}$ (see Section 4.3).
4. *Baseline* + Set_{all}: The NMT model is trained on the original parallel corpus along with phrase pair set Set_{all} (see Section 4.3).

The number of training examples for the systems as mentioned above are shown in Table 3.

7. Results and analysis

We evaluate the models on the test sets using BLEU metric (Papineni *et al.* 2002). Table 4 summaries the results of different systems, and in Tables 9 and 10, we show some example outputs obtained from these systems. We plot the BLEU scores of the different translation systems in Figures 4 and 5 for comparison with the baselines.

7.1 Attention-based GRU versus transformer

Both attention-based GRU and transformer-based models are improved by our phrase-augmentation approach. However, if we compare them, transformer-based models are better than the attention-based GRU models for all the translation directions. Transformer-based models result in better baselines than the GRU-based models. Also, SMT is known to work better than the neural models in the absence of sufficient training data. But with our approach, transformer-based models perform better than the SMT-based baselines for five translation directions (see Table 4), and for rest of the translation directions, our approach with transformer networks obtains the competitive (with SMT) results.

7.2 Comparative systems

Here, we compare our proposed approach with some of the well-explored techniques in low-resource NMT such as subword-level NMT (Sennrich, Haddow, and Birch 2016b), back

[b]Sockeye supports word-based batching too.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 4. BLEU scores of different models

	Hindi→English			English→Hindi			Hindi→Bengali		Bengali→Hindi		Old→Mod
Model	H	T	J	H	T	J	H	T	H	T	English
SMT	23.07	24.39	29.36	20.64	19.24	26.75	28.50	25.09	29.81	29.62	39.95
Attention-based GRU											
Baseline	14.02	12.14	6.97	14.04	12.26	11.25	18.11	13.81	21.2	19.04	10.03
+ $Set_{p\geq 0.5}$	18.53	**20.79**	18.21	14.88	15.95	**20.69**	20.39	**19.25**	21.64	20.46	25.41
+ $Set_{p=1.0}$	**19.43**	19.75	**22.33**	**16.35**	**16.67**	17.19	20.70	18.76	**22.58**	**22.06**	20.83
+ Set_{all}	18.73	18.22	21.39	16.14	14.92	19.74	**20.93**	18.08	21.45	21.42	**28.76**
▲	5.41 ↑	8.65 ↑	15.36 ↑	2.31 ↑	4.41 ↑	9.44 ↑	2.82 ↑	5.44 ↑	1.38 ↑	3.02 ↑	18.73 ↑
Transformer											
Baseline	22.08	20.41	25.34	20.71	17.44	21.84	22.81	18.90	24.92	22.92	27.94
+ $Set_{p\geq 0.5}$	**25.70***	23.59	**29.85***	22.24	19.66	**25.99**	**27.04**	**24.00**	28.74	26.42	33.40
+ $Set_{p=1.0}$	24.65	**24.40***	29.43	**23.97***	**19.74***	25.50	26.17	23.47	**28.90**	**26.88**	**32.67**
+ Set_{all}	25.63	23.80	29.19	23.01	18.19	25.53	26.03	23.73	28.51	26.86	33.61
▲	3.68 ↑	3.99 ↑	4.51 ↑	3.26 ↑	2.30 ↑	4.15 ↑	4.23 ↑	5.10 ↑	3.98 ↑	3.96 ↑	5.67 ↑

▲, improvement over baseline; ↑, positive improvement; *, Better than SMT; H, Health; T, Tourism; J, Judicial. Highest BLEU score for each translation direction using NMT indicated in bold.

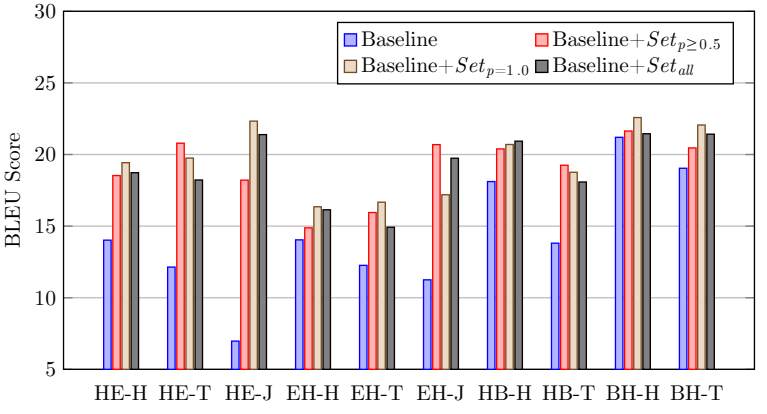

Figure 4. Comparison of different attention-based GRU models. HE: Hindi→English, EH: English→Hindi, HB: Hindi→Bengali, BH: Bengali→Hindi. H for Health, T for Tourism and J for Judicial.

translation (Sennrich *et al.* 2016a), and more in line with our proposed approach, pre-translation (Niehues *et al.* 2016). Recently, transformer-based models have outperformed the attention-based GRU models on the various benchmark datasets and have become the state-of-the-art technique in NMT. This is also evident from the evaluation results that we obtain. Hence, we focus only on the transformer-based models for comparison. We compare our approach for English–Hindi translation direction involving Health domain data.

Pre-translation for NMT Niehues *et al.* (2016) have proposed two methods to improve the NMT with the help of phrase-based SMT system. In the first method, Niehues *et al.* (2016) first

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 5. Comparative systems for English–Hindi for Health domain

System	BLEU score
SMT	20.64
Transformer	20.71
Pre-translation (Niehues *et al*. 2016)	20.40
Mixed pre-translation (Niehues *et al*. 2016)	19.58
Transformer with back translation	18.75
Proposed best model	23.97

Table 6. Comparison of our approach (using attention-based GRU) with PhraseNet for English–Hindi for Health domain. In parenthesis, we show the dimension

System	BLEU score
PhraseNet (embedding = 620 and hidden = 1000)	9.62
PhraseNet (embedding = 128 and hidden = 256)	12.05
PhraseNet (embedding = 300 and hidden = 600)	12.65
Our approach with $Set_{p\geq 0.5}$	15.95
Our approach with $Set_{p=1.0}$	16.67
Our approach with Set_{all}	14.92

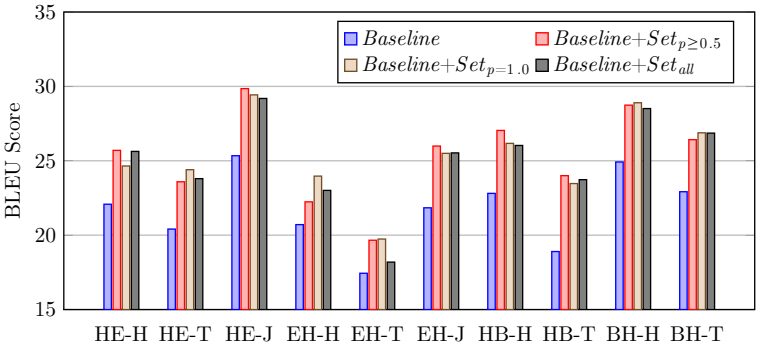

Figure 5. Comparison of different transformer models. HE: Hindi→English, EH: English→Hindi, HB: Hindi→Bengali, BH: Bengali→Hindi. H for Health, T for Tourism, and J for Judicial.

trained a source–target SMT system, and then using the SMT system, they translated the entire training data from source to target. Thereafter, they trained a monolingual NMT system from the translated-target to original target. Second method is almost the same as the first but the NMT system is trained to predict the target from the combination of original source and translated output of SMT system. The difference between our approach and the pre-translation technique is that we consider the phrase pairs extracted from the phrase table as the additional parallel data, whereas the authors (Niehues *et al*. 2016) trained a monolingual NMT system to correct the outputs produced by a phrase-based SMT system. We follow the same approach as in Niehues *et al*. (2016), and the results are shown in Table 5.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 7. Our approach using transformer at subword level for English–Hindi direction on the Health domain

System	BLEU score
Subword (Sennrich *et al.* 2016b)	21.58
Our Approach with $Set_{p=0.5}$	23.27
Our Approach with $Set_{p=1.0}$	22.33
Our Approach with Set_{all}	23.18

From Table 5, we observe that pre-translation and mixed pre-translation strategies do not improve the SMT and NMT models but, rather, they degrade their baselines.

Back-translation For this, we first generate synthetic parallel data by translating 100K monolingual sentences from the Hindi monolingual data (Bojar *et al.* 2014) into English. Then, we use these synthetic parallel sentences along with the original parallel data to train a transformer-based system for English→Hindi. From Table 5, we observe that the BLEU score is lower than that of the system (*Transformer*) using only the original parallel data. Although it has been shown in the literature that back translation helps in improving the BLEU score, but it is also sensitive to the domain of the back-translated data. The monolingual Hindi data (Bojar *et al.* 2014) are from a mixed domain, crawled from the web. This shows that back translation may not be always useful.

PhraseNet Tang *et al.* (2016) have proposed *PhraseNet* in which decoder generates a word in *word mode* or a phrase in *phrase mode*. As the code of the PhraseNet is not available, we re-implement it to compare with our proposed method. Here, we first mathematically describe the approach and then present our results. Suppose, at time $t-1$, the decoder has generated y_{t-1} in word mode and the current decoder state is s_t.

1. Compute the *word mode* $(=1)$ and *phrase mode* $(=0)$ probabilities as
$$p(z_t = 1|s_t;\theta) = f_z(s_t)$$
$$p(z_t = 0|s_t;\theta) = 1 - f_z(s_t)$$
2. If $z_t = 0$ (*word mode*), generate a word w_i based on s_t from the regular word vocabulary as
$$p_w(y_t = w_i|s_t, 0;\theta) = f_w(s_t)$$
3. If $z_t = 1$ (*phrase mode*), generate target phrase p_j as
$$p_p(y_t = p_j|s_t, 1;\theta) = f_p(s_t)$$
4. Calculate the final probabilities and sample the next word (or phrase):
$$p(y_t = w_i) = p(z_t = 0|s_t;\theta)p(w_i|s_t, 0;\theta)$$
$$p(y_t = p_j) = p(z_t = 1|s_2;\theta)p(p_j|s_t, 1;\theta)$$
$$p(y_t) = \begin{bmatrix} p(y_t = w) \\ p(y_t = p) \end{bmatrix}$$
where the size of $p(y_t)$ is n_p plus the number of words in the vocabulary. The value of the hyper-parameter n_p is set to 5. The next word or phrase will be sampled according to $p(y_t)$. For more details, see Tang *et al.* (2016).

We reimplement PhraseNet using PyTorch (Paszke *et al.* 2017). We compare our approach with *PhraseNet* for English–Hindi translation direction involving Health domain data. Tang *et al.* (2016) have experimented with embedding dimension 620 and hidden dimension 1000. As our

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 8. Fluency: SMT versus NMT systems. Word-to-word translation shown in brackets

Source	Cough fills up inside lungs in the second stage.
Reference	दूसरी अवस्था में फेफड़ों में कफ भर जाता है ।
	(second stage in lungs inside cough fills up)
SMT	फेफड़ों में कफ भर दूसरी स्टेज में है ।
	(lungs in cough fills second stage in)
NMT	दूसरे चरण में फेफड़ों के अंदर कफ भर जाता है ।
	(second stage in lungs inside cough fills up.)

data size is smaller, we experiment with different embedding dimensions and hidden dimensions. We present the results in Table 6. We observe from this table that our proposed approach outperforms PhraseNet with a significant BLEU point.

Subword-level NMT We train subword-level transformer systems (baseline and our approach) for English–Hindi direction for Health domain, and the results are shown in Table 7. We consider 10,000 merge operations for each language independently. From Table 7, we see that our approach at the subword level also outperforms the baseline system.

7.3 Quantitative analysis

From Figure 4, it is very much evident that the baseline NMT systems are outperformed by all of our proposed systems based on attention-based GRU and transformer. As the data size is too small for an NMT, we feed the phrase-level translation during training. The intuition is that the added phrases provide more information about the association among the phrases, and this helps in the learning process. The baseline model finds it too difficult to learn this association from the original training set with a few parallel sentences. The baselines for different language pairs (and translation directions) are trained on the original training sentences, whereas in our proposed method, we feed the extracted phrases into the model. We can see the differences those additional phrase-level translations make. Although we do not use any external data, still, we obtain significant improvements over the baselines for all the translation systems. We observe the improvement of 1.38 for *Bengali→ Hindi(Health)* to 8.65 for *Hindi→ English (Tourism)* in BLEU points on ILCI data. For *Hindi→English (Tourism)*, we observe the highest improvement.

Surprisingly, for all the translation directions, the improvements are more in case of the tourism domain compared to the health domain. The possible reason behind this can be explained as follows: tourism domain corpora have more named entities (mostly location names) compared to the health domain corpora, and when we feed phrase pairs as sentences, those named entities (source–target pairs) are also included in the system. Thus, the system learns better alignment between these, and as a result, the overall translation quality is improved. We see, from the table, that without these additional phrases the baseline for tourism domain always performs poorer than the baselines for the health domain. Thus, the additional phrases make the improvements more visible.

We also notice that when translating from morphologically rich (Hindi) to poor (English) language, the improvements are higher as compared to the setup in the opposite direction, that is, English to Hindi translation.

For old-to-modern English system, the baseline model has a BLEU of 10.03 and the proposed models are better than the baseline model. Out of the three proposed models, the model using all phrases (*Baseline+*Set_{all}) yields the best performance with 28.76 BLEU points. However, the difference between the baseline and the proposed models is huge because the old-to-modern systems

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 9. Some translation outputs by different attention-based encoder–decoder NMT systems. Word-to-word English translation is shown in brackets

	English → Hindi (Health)
Source	Cough fills up inside lungs in the second stage.
Reference	दूसरी अवस्था में फेफड़ों में कफ भर जाता है । (second stage in lungs inside Cough fills up)
Baseline	दूसरी स्टेज में दूसरी स्थिति में श्लेष्मा भर जाता है । (second stage in second stage in mucus fills up)
$+Set_{p\geq 0.5}$	खाँसी के रूप में कफ को एकत्रित होता है । (cough of form in cough collected is.)
$+Set_{p=1.0}$	दूसरे चरण में कफ को अन्दर चला जाता है । (second stage in cough inside goes.)
$+Set_{all}$	दूसरे चरण में फेफड़ों के अंदर कफ भर जाता है । (second stage in lungs inside cough fills up)
	English → Hindi (Tourism)
Source	The idol of the God is made of black stone.
Reference	भगवान की मूर्ति काले पत्थर से निर्मित है । (God of idol black stone of made is.)
Baseline	देवता की प्रतिमा काले पत्थर की बनी है । (God of idol black stone of made is.)
$+Set_{p\geq 0.5}$	देवता की मूर्ति काले पत्थर की बनी है । (God of idol black stone of made is.)
$+Set_{p=1.0}$	देवता की मूर्ति काले पत्थर की बनी है । (God of idol black stone of made is.)
$+Set_{all}$	भगवान मूर्ति की प्रतिमा काले पत्थर से बनी है । (God idol of idol black stone of made is.)
	Hindi → Bengali (Health)
Source	30 डिग्री से कम तापमान होने पर त्वचा नीली पड़ जाती है ।
Translation	If the temperature is less than 30 degrees, the skin becomes blue.
Reference	তাপমাত্রা ৩০ ডিগ্রির কম হওয়ার ফলে ত্বক নীল হয়ে যায়। (temperature 30 degree less is due skin blue turns.)
Baseline	৩০ ডিগ্রি থেকে কম তাপমাত্রা লাগার ফলে ত্বক সেরে যায়। (30 degree than less temperature feel due skin cured gets.)
$+Set_{p\geq 0.5}$	৩০ ডিগ্রিরও কম তাপমাত্রা হলে ত্বক নীল হয়ে যায়। (30 degree less temperature is skin blue turns.)
$+Set_{p=1.0}$	৩০ ডিগ্রিরও কম তাপমাত্রা হওয়ার ফলে ত্বক নীল হয়ে যায়। (30 degree less temperature is due skin blue turns.)
$+Set_{all}$	৩০ ডিগ্রি থেকে কম তাপমাত্রা হলে ত্বক নীল হয়ে যায়। (30 degree than less temperature is skin blue becomes.)
	Hindi → Bengali (Tourism)
Source	सफदरजंग रोड पर इंदिरा गाँधी मेमोरियल है , जहाँ उनकी हत्या हुई थी ।
Translation	Indira Gandhi Memorial at Safdarjung Road, where she was killed.
Reference	সফদরগঞ্জ রোডের উপর ইন্দিরা গান্ধী মেমোরিয়াল আছে , যেখানে ওনার হত্যা হয়েছিল। (Safdarjung road on Indira Gandhi memorial is, where her murder was.)
Baseline	মাঝীর অভয়ারণ্যে ইন্দিরা গান্ধী গোলাকার ভবন আছে যেখানে ওনার প্রতিমাকে এখানে প্রতিষ্ঠিত হয়। (Majhir sanctuary in Indira Gandhi rounded building is where her idol here established was.)
$+Set_{p\geq 0.5}$	সফদরগঞ্জের রোডে ইন্দিরা গান্ধী মেমোরিয়াল আছে , যেখানে তার উতসব ছিল। (Safdarjung on road Indira Gandhi memorial is, where her festival was.)
$+Set_{p=1.0}$	সফরদরজঙ্গ রোডে ইন্দিরা গান্ধী মেমোরিয়াল আছে , যেখানে তার মহিমা প্রচার হয়। (Safdarjung on road Indira Gandhi memorial is,where her greatness promoted is.)
$+Set_{all}$	সফরদরজঙ্গ রোডে ইন্দিরা গান্ধী মেমোরিয়াল আছে , যেখানে ওনার প্রভাবিত হয়েছিল। (Safdarjung on road Indira Gandhi memorial is, where her affected was.)

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 9. Continued

	Hindi → Bengali (Tourism)
Source	चंडीगढ़ में आप नेकचंद के रॉकगार्डन के अलावा रोज गार्डन , सुखना लेक तथा यूनिवर्सिटी की खूबसूरती का नजारा ले सकते हैं ।
Translation	In Chandigarh, you can take a look at the beauty of Rose Garden, Shushana Lake and University besides the Rock Garden of Nekchand.
Reference	চণ্ডীগড়ে আপনি নেকচন্দের রকগার্ডেন ছাড়াও রোজ গার্ডেন , সুখনা লেক , এবং ইউনিভার্সিটির সৌন্দর্য উপভোগ করতে পারবেন। (Chandigarh you Nekchand's RockGardens besides rose garden, Sukhna lake, and university's beauty enjoy can.)
Baseline	ঘোরার মত ছাদওয়ালা পাওয়া যাবে ভিড়ভাড় , সুখনা লেক এবং কক্ষের সৌন্দর্যকে আনন্দ উপভোগ করতে পারবেন। (Roam to rooftop found be rush, Sukhna lake and room's beauty enjoy can.)
$+Set_{p\geq 0.5}$	গরমের সময়ে স্নান করার জন্য রোজ গার্ডেন , সুখনা লেক এবং ইউনিভার্সিটি বলে মনে করতে পারেন। (summer during bathing for rose garden, Sukhna lake and university to be think.)
$+Set_{p=1.0}$	কোলকাতায় আপনি বারমের পার্ক ছাড়াও রোজ গার্ডেন , সুখনা লেক এবং পাণ্ডার সৌন্দর্যের দৃশ্য উপভোগ করতে পারেন। (in Kolkata you Barmer park besides rose garden and panda's beauty view enjoy can.)
$+Set_{all}$	দীঘায় আপনি নেকচান্দ ছাড়াও রোজ গার্ডেন , সুখনা এবং ইউনিভার্সিটি অনেক সৌন্দর্যের দৃশ্য উপভোগ করতে পারেন। (in Digha you Nekchand besides rose garden, Sukhna and university many beauty view enjoy can.)

have original training data of only 2.7K parallel sentences. As a result, the baseline model does not learn the mappings well and our models have better scope in learning the mappings.

Along with NMT, we also train SMT systems. SMT systems are known to be good for situations when we do not have enough parallel corpus, and with no surprise, we observe that SMT models perform better than the NMT models. However, there are good reasons for considering NMT when we do not have sufficient amount of parallel corpus. Improvement of NMT quality with the increase in data size is huge as compared to SMT (Koehn and Knowles 2017). Other reasons to consider SMT are NMT follows an end-to-end framework and generates more fluent outputs than the SMT systems.

The phrases that are added to the original training corpus have lengths from 1 to 7. So, we did a study on the effect of phrase length on the overall system performance. Apart from the sets as mentioned in Section 4.3, we also consider only the phrases with lengths 1, 3, 5, and 7 from the set $Set_{p=1.0}$ (English→Hindi; tourism) for augmenting to the original training corpus. Experiments show that all these sets of phrases improve the performance of the baseline system. However, while we consider all the phrases (of length 1–7), we obtain the best BLEU score. Experimental results are shown in Figure 7.

We use three types of training data in terms of the number of sentences: with 2.7K parallel sentences (for old-to-modern English), 5.5k parallel sentences (for Judicial domain), and 23k parallel sentences (for Health and Tourism domains). However, the improvements are higher for the smaller datasets (c.f. Table 1 for data size and Table 4 for the improvements). For example, old-to-modern English system has the highest improvement, whereas the systems trained using 23k sentences have relatively less improvements. Hence, we take one system (English→Hindi; tourism) with lower improvement and apply our proposed approach to see how it behaves while we have relatively smaller sets (of 5k, 10k, and 15k parallel sentences). These smaller sets were taken from the original training data.

From the experimental results as shown in Figure 6, we observe that improvements are higher for the smaller amount of data. This implies that NMT models, as we know, fail when the data size is small, but the extracted phrases help up to a certain level.

Since we use phrase pairs as training data, it is obvious that most of the training samples are relatively short. Thus, it is interesting to see if this affects the translation quality for different

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

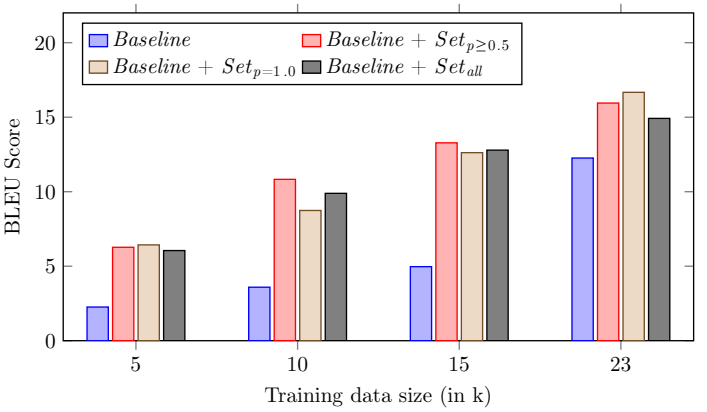

Figure 6. Comparison of different NMT models with incremental original training data.

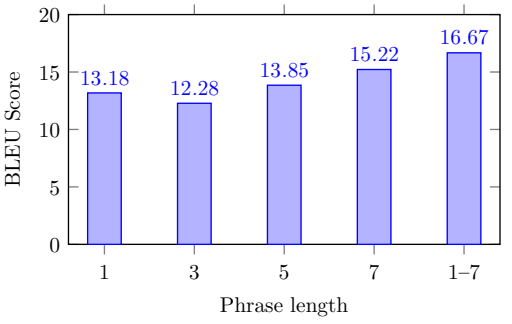

Figure 7. Performance of proposed method with different phrase lengths.

length intervals. For this, we split the testset for English–Hindi (Health) according to the different length intervals (such as <10, 10–20, 20–30, 30–40, and >40) and score using transformer ($set_{p=1}$)-based model. The BLEU scores for these intervals are 23.84, 26.67, 21.03, 21.23, and 27.63, and the sentence counts are 172, 508, 236, 60, and 18, respectively.

7.4 Qualitative analysis

In this section, we present our observations on the quality of outputs produced by the proposed systems as compared to the baseline models. From Table 9, for *English→Hindi (Health)*, we observe that the output of the baseline system is not adequate as the translation of "lungs" is dropped and "second stage" is translated twice. These kinds of errors are very common in any NMT system. This is because the baseline model (trained only on the original training set) does not learn the mappings between these phrases as the corpus is very small. In contrast, our proposed system produces better translation output.

From Table 9, we observe that for English → Hindi (Tourism) system, all the models including baseline generate the good-quality translations. However, there are some failed cases. For example, the proposed model *Baseline+*Set_{all} over-translates the source word "idol."

Now we look into the quality of translation produced by the related language pair, for example, Hindi–Bengali. The Hindi → Bengali (Health) baseline system generates the incorrect output, and the outputs generated by our proposed systems are of good quality.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Table 10. Some translation outputs by different transformer-based models. Word-to-word English translation is shown in brackets

	Hindi → Bengali (Tourism)
Source	सफदरजंग रोड पर इंदिरा गाँधी मेमोरियल है , जहाँ उनकी हत्या हुई थी ।
Translation	Indira Gandhi Memorial at Safdarjung Road, where she was murdered.
Reference	সফদরগঞ্জ রোডের উপর ইন্দিরা গান্ধী মেমোরিয়াল আছে , যেখানে ওনার হত্যা হয়েছিল।
	(Safdarjung road on Indira Gandhi memorial is, where her murder was.)
Baseline	লাকোট রোডের উপর ইন্দিরা গান্ধী মেমোরিয়াল আছে যেখানে তাঁর দিব্যশক্তিযুক্ত লীলা ছিল।
	Lakot road on Indira Gandhi memorial is, where her divine power play was.)
$+Set_{p\geq 0.5}$	সফরদরজঙ্গ রোডের উপর অবস্থিত ইন্দিরা গান্ধী মেমোরিয়াল আছে যেখানে ওনার পরমজ্ঞানের উপলব্ধি হয়েছিল।
	(Safdarjung road on Indira Gandhi memorial is, where her perfection realized was.)
$+Set_{p=1}$	সফরদরজঙ্গ রোডে ইন্দিরা গান্ধী মেমোরিয়াল আছে , যেখানে ওনার হত্যা করেন।
	(Safdarjung road on Indira Gandhi memorial is, where she murdered was.)
$+Set_{all}$	সফরদরজঙ্গ রোডে ইন্দিরা গান্ধী মেমোরিয়াল আছে যেখানে ওনার হত্যা করা হয়।
	(Safdarjung road on Indira Gandhi memorial is where she murdered was.)

For Hindi → Bengali (Tourism), the output of the baseline system is incorrect and only a few source words (इंदिरा गाँधी, है, जहाँ, उनकी) are correctly translated, whereas our proposed systems produce partially correct translations. However, one important point we observe for Hindi → Bengali (Tourism) models: all the three systems made similar kinds of mistakes by incorrectly translating the place names. For example, while the baseline model drops the translation of "चंडीगढ़" (Chandigarh), our two proposed models wrongly translate the place names into "কোলকাতা" (Kolkata) and "দীঘায়" (Digha). One advantage of continuous representation of words is that it enables NMT to learn the semantic similarity between the related words (e.g., house and home). However, this also introduces a drawback in the NMT system, which often wrongly translates into the words that seem natural in the context but does not reflect the source words.

We also conduct a study on the fluency of output translations. Although the SMT systems have better BLEU scores than the NMT systems, we found that NMT outputs are more fluent than the SMT ones. Few examples are shown in Table 8. From Table 4, we see the Bengali→Hindi for Health domain has the least improvement (1.38 BLEU points) compared to that of the other systems. In order to check if the said improvement is significant, we perform the statistical significance test based on the bootstrap re-sampling (Koehn 2004). We found that the improvement is significant at the confidence level 95% with p-value 0.001.

8. Conclusion and future works

In this paper, we have proposed an approach for training an NMT model using a small parallel corpus. Our approach uses the phrase pairs extracted from the original training corpus as feedback to NMT training. We followed the attention-based GRU and transformer architectures for experiments. However, the proposed approach is not specific to these architecture only. It can also be applied to the other NMT architectures as well. We used publicly available English–Hindi and Hindi–Bengali parallel corpora in three and two domains, respectively, for the evaluation. We also applied our proposed approach to an interesting translation task that focused on old-to-modern English translation. We found that the proposed method significantly improves over

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

the baseline system when we have far from sufficient amount of parallel corpus. Improvement in BLEU is approximately 1.38–15.36 points. For old-to-modern English translation, we observed significant improvement of more than 18 BLEU points over the baseline model.

In future, our main focus will study the effect of phrase augmentation for language pairs with bigger (sufficient) corpora—if extracted phrases help or they are just redundant.

Acknowledgments. Asif Ekbal gratefully acknowledges Young Faculty Research Fellowship, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Artetxe M., Labaka G. and Agirre E. (2018). Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 3632–3642.

Arthur P., Neubig G. and Nakamura S. (2016). Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1557–1567.

Bahdanau D., Cho K. and Bengio Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representation (ICLR)*.

Bojar O., Chatterjee R., Federmann C., Graham Y., Haddow B., Huck M., Yepes A.J., Koehn P., Logacheva V., Monz C., Negri M., Névéol A., Neves M., Popel M., Post M., Rubino R., Scarton C., Specia L., Turchi M., Verspoor K. and Zampieri M., (2016). Findings of the 2016 conference on machine translation. In *ACL 2016 First Conference on Machine Translation (WMT16)*. The Association for Computational Linguistics, pp. 131–198.

Bojar O., Diatka V., Rychlỳ P., Stranák P., Suchomel V., Tamchyna A. and Zeman D. (2014). Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC*, pp. 3550–3555.

Cho K., Van Merriënboer B., Bahdanau D. and Bengio Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111.

Crego J., Kim J., Klein G., Rebollo A., Yang K., Senellart J., Akhanov E., Brunelle P., Coquard A., Deng Y., Enoue S., Geiss C., Johanson J., Khalsa A., Khiari R., Ko B., Kobus C., Lorieux J., Martins L., Nguyen D.-C., Priori A., Riccardi T., Segal N., Servan C., Tiquet C., Wang B., Yang J., Zhang D., Zhou J. and Zoldan P. et al. (2016). Systran's pure neural machine translation systems. arXiv preprint arXiv:1610.05540.

Fadaee M., Bisazza A. and Monz C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 567–573.

Feng Y., Zhang S., Zhang A., Wang D. and Abel A. (2017). Memory-augmented neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1390–1399.

Forcada M.L. and Ñeco R.P. (1997). Recursive hetero-associative memories for translation. In *International Work-Conference on Artificial Neural Networks*. Springer, pp. 453–462.

Gulcehre C., Firat O., Xu K., Cho K. and Bengio Y. (2017). On integrating a language model into neural machine translation. *Computer Speech & Language* **45**, 137–148.

Guzmán F., Chen P.-J., Ott M., Pino J., Lample G., Koehn P., Chaudhary V. and Ranzato M. (2019). Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. arXiv preprint arXiv:1902.01382.

He W., He Z., Wu H. and Wang H. (2016). Improved neural machine translation with smt features. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Heafield K. (2011). Kenlm: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pp. 187–197.

Hieber F., Domhan T., Denkowski M., Vilar D., Sokolov A., Clifton A. and Post M. (2017). Sockeye: a toolkit for neural machine translation. arXiv preprint arXiv:1712.05690.

Hochreiter S. and Schmidhuber J. (1997). Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

Jha G.N. (2010). The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

Junczys-Dowmunt M., Dwojak T. and Hoang H. (2016). Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Kalchbrenner N. and Blunsom P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709.

Kingma D.P. and Ba J. (2015). Adam: a method for stochastic optimization. In *International Conference on Learning Representation (ICLR)*.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Kneser R. and Ney H. (1995). Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95*, vol. 1. IEEE, pp. 181–184.

Koehn P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

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. and Herbst E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, pp. 177–180.

Koehn P. and Knowles R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver. Association for Computational Linguistics, pp. 28–39.

Koehn P., Och F.J. and Marcu D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pp. 48–54.

Kunchukuttan A., Mehta P. and Bhattacharyya P. (2018). The IIT Bombay English-Hindi parallel corpus. In Calzolari N., Choukri K., Cieri C., Declerck T., Goggi S., Hasida K., Isahara H., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., Piperidis S. and Tokunaga T. (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Lample G. and Conneau A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

Lample G., Ott M., Conneau A., Denoyer L. and Ranzato M. (2018). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 5039–5049.

Niehues J., Cho E., Ha T.-L. and Waibel A. (2016). Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan. The COLING 2016 Organizing Committee, pp. 1828–1836.

Papineni K., Roukos S., Ward T. and 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*, Philadelphia, Pennsylvania, pp. 311–318.

Paszke A., Gross S., Chintala S., Chanan G., Yang E., DeVito Z., Lin Z., Desmaison A., Antiga L. and Lerer A. (2017). Automatic differentiation in pytorch. In *NIPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*.

Ramachandran P., Liu P. and Le Q. (2017). Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 383–391.

Ren S., Zhang Z., Liu S., Zhou M. and Ma S. (2019). Unsupervised neural machine translation with smt as posterior regularization. arXiv preprint arXiv:1901.04112.

Sen S., Hasanuzzaman M., Ekbal A., Bhattacharyya P. and Way A. (in press). Take help from elder brother: old to modern english nmt with phrase pair feedback. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*.

Sennrich R., Firat O., Cho K., Birch A., Haddow B., Hitschler J., Junczys-Dowmunt M., Läubli S., Miceli Barone A.V., Mokry J. and Nadejde M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 65–68.

Sennrich R., Haddow B. and Birch A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany*.

Sennrich R., Haddow B. and Birch A. (2016b). 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)*, Berlin, Germany. Association for Computational Linguistics, pp. 1715–1725.

Song K., Zhang Y., Yu H., Luo W., Wang K. and Zhang M. (2019). Code-switching for enhancing NMT with pre-specified 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)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 449–459.

Sutskever I., Vinyals O. and Le Q.V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112.

Tang Y., Meng F., Lu Z., Li H. and Yu P.L. (2016). Neural machine translation with external phrase memory. arXiv preprint arXiv:1606.01792.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł. and Polosukhin I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008.

https://doi.org/10.1017/S1351324920000303 Published online by Cambridge University Press

Wang X., Lu Z., Tu Z., Li H., Xiong D. and Zhang M. (2017). Neural machine translation advised by statistical machine translation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Wang X., Pham H., Dai Z. and Neubig G. (2018). SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 856–861.

Wang X., Tu Z. and Zhang M. (2018). Incorporating statistical machine translation word knowledge into neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **26**(12), 2255–2266.

Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Liu X., Kaiser U., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M. and Dean J. (2016). Google's neural machine translation system: bridging the gap between human and machine translation. CoRR abs/1609.08144.

Zhang J. and Zong C. (2016). Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545.

Zhang Z., Liu S., Li M., Zhou M. and Chen E. (2018). Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhao Y., Wang Y., Zhang J. and Zong C. (2018). Phrase table as recommendation memory for neural machine translation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13–19, 2018, Stockholm, Sweden*, pp. 4609–4615.

Zoph B., Yuret D., May J. and Knight K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pp. 1568–1575.

Cite this article: Sen S, Hasanuzzaman M, Ekbal A, Bhattacharyya P and Way A (2021). Neural machine translation of low-resource languages using SMT phrase pair injection. *Natural Language Engineering* **27**, 271–292. https://doi.org/10.1017/S1351324920000303