Machine Translation (2021) 35:661–685
https://doi.org/10.1007/s10590-021-09290-0

Augmenting training data with syntactic phrasal-segments in low-resource neural machine translation

Kamal Kumar Gupta[1] · Sukanta Sen[1] · Rejwanul Haque[2] · Asif Ekbal[1] · Pushpak Bhattacharyya[1] · Andy Way[2]

Received: 25 February 2020 / Accepted: 12 November 2021 / Published online: 21 December 2021

Abstract

Neural machine translation (NMT) has emerged as a preferred alternative to the previous mainstream statistical machine translation (SMT) approaches largely due to its ability to produce better translations. The NMT training is often characterized as *data hungry* since a lot of training data, in the order of a few million parallel sentences, is generally required. This is indeed a bottleneck for the under-resourced languages that lack the availability of such resources. The researchers in machine translation (MT) have tried to solve the problem of data sparsity by augmenting the training data using different strategies. In this paper, we propose a generalized linguistically motivated data augmentation approach for NMT taking low-resource translation into consideration. The proposed method operates by generating source—target phrasal segments from an authentic parallel corpus, whose target counterparts are linguistic phrases extracted from the syntactic parse trees of the target-side sentences. We augment the authentic training corpus with the parser generated phrasal-segments, and investigate the efficacy of our proposed strategy in low-resource scenarios. To this end, we carried out experiments with resource-poor language pairs, *viz.* Hindi-to-English, Malayalam-to-English, and Telugu-to-English, considering the three state-of-the-art NMT paradigms, *viz.* attention-based recurrent neural network (Bahdanau et al., 2015), Google Transformer (Vaswani et al. 2017) and convolution sequence-to-sequence (Gehring et al. 2017) neural network models. The MT systems built on the training data prepared with our data augmentation strategy significantly surpassed the state-of-the-art NMT systems with large margins in all three translation tasks. Further, we tested our approach along with back-translation (Sennrich et al. 2016a), and found these to be complementary to each other. This joint approach has turned out to be the best-performing one in our low-resource experimental settings.

Keywords Neural machine translation · Low-resource neural machine translation · Data augmentation · Syntactic phrase augmentation

Extended author information available on the last page of the article

1 Introduction

Neural Machine Translation (NMT) (Bahdanau et al. 2015; Vaswani et al. 2017; Gehring et al. 2017) is an end-to-end learning approach in the automatic translation where a deep neural network (NN) is trained by deep learning techniques. Over the past 6 years, NMT has steadily superseded the previous mainstream machine translation (MT) techniques e.g. Phrase-based Statistical Machine Translation (PB-SMT) (Koehn et al. 2003), both in academia and industry, and now, it is a new state-of-the-art in MT research. In recent years, NMT has reached its new heights in field of MT, where even some researchers claim to have parity with human translation (Hassan et al. 2018). Despite its huge successes, NMT has a number of weaknesses, e.g. NMT is found to be under-performing for low-resource language-pairs (Koehn and Knowles 2017). This is indeed problematic for the language pairs which do not have sufficient parallel samples for training the NMT model.

Machine translation researchers have tried to counter this problem by introducing various methods, e.g. using source and target side monolingual data in training, augmenting parallel training data, exploiting training data involving other language pairs. Back-translation (Sennrich et al. 2016b) is by far viewed as the most successful strategy (Edunov et al. 2018) for data augmentation in NMT, and is regarded as the most popular method to the MT developers since it is a relatively simple process and can greatly minimise the data sparseness problem using the target monolingual data only (Edunov et al. 2018; Wang et al. 2019).

In this paper, we present a novel data augmentation method for NMT that generates additional training examples (i.e. phrasal segment-pairs) from an authentic parallel training corpus. In short, we at first extract three types of syntactic phrases, namely noun phrases (NPs), verb phrases (VPs) and prepositional phrases (PPs), from the constituency trees of the target language sentences of the parallel training corpus, obtain the source translations of the extracted phrases by back-translating them using a target-to-source MT system, and create an additional synthetic corpus with the resulting linguistically coherent source-target phrasal segments. Thereafter, we augment the original training data by adding this additional synthetic corpus. We carried out our experiments with three less-examined and low-resource language-pairs, *viz.* Hindi-to-English, Malayalam-to-English and Telugu-to-English, and considered three state-of-the-art NMT approaches, namely the attention-based recurrent neural network (ARNN) (Bahdanau et al. 2015), Google Transformer (GT) (Vaswani et al. 2017) and convolution sequence-to-sequence neural network (CSSNN) (Gehring et al. 2017) models, as our baselines. The MT systems, which were trained following our strategy statistically significantly outperform the state-of-the-art NMT baselines in all the translation tasks.

Furthermore, we compare our approach with the state-of-the-art back-translation approach (Sennrich et al. 2016b) in this low-resource translation setting. We also applied our proposed and the back-translation approaches together to see how they would perform in a resource-poor scenario. They were found to be complementary to each other, and the MT systems that were built following this collaborative strategy produced the best BLEU (Papineni et al. 2002) gains in all the three translation tasks.

The remainder of the paper is organized as follows. Section 2 presents a brief survey of the related work. Section 4 describes our linguistically-motivated data augmentation method. Section 5 presents the details of our datasets, and Sect. 7 explains the experimental setup. In Sect. 9, we present the results along with appropriate discussions and analysis. Finally, Sect. 10 concludes the work along with the future work.

2 Related work

In this section we present a survey on the prior research that are very closely related to our current work. Sennrich et al. (2016b) introduced the back-translation based method. They translated target side monolingual data to create a synthetic parallel corpus and merged it to the original parallel corpus. Augmented parallel data (original parallel data along with synthetic parallel data) was then used for training the model. Inspired by the back-translation technique, Zhang and Zong (2016) proposed a method by adding the translated monolingual source-side data to the target-side, and creating the synthetic parallel data. Zoph et al. (2016) used the transfer learning method by initializing parameters of the low-resource language pair with parameters of a high resource language pair model. This approach in general performs better when both the language pairs are related. In case of being distant or having different domains for high and low resource languages pairs, transfer learning may not show much improvement.

Currey et al. (2017) introduced a copied parallel corpus in which an identical copy of the target side monolingual data is added at the source side which was used along with the original training data for training NMT models. Fadaee et al. (2017) generated synthetic training data by augmenting rare vocabulary words into the existing sentences. From vocabulary, they chose a word as rare if its frequency is less than some threshold value. Then by using a language model they replaced the rare words in the sentences. Wang et al. (2018) replaced words in both the source and target sentences randomly with the other words in vocabulary by using some sampling techniques. They showed the improvement over the baselines using this augmentation technique. Fadaee and Monz (2018) also used the back-translation technique for creating synthetic data. But instead of randomly selecting monolingual sentences they suggested selecting the sentences which have words with high prediction loss or which are difficult to predict. Zhu et al. (2019) introduced an augmentation technique by replacing a token in a sentence with a contextual combination of its related words. They replaced the embedding of a word by a weighted combination of multiple words which are semantically similar to that word.

Zhao et al. (2018) used the phrase table as a recommendation memory for adding a bonus score to the words which are more adequate and worthy of recommendation. Wang et al. (2017) used phrases from the phrase table generated using SMT. These phrases are used to write a phrase memory that is used during the decoding in NMT. Aharoni and Goldberg (2017) introduced a string-to-tree approach in order to improve the translation performance in NMT. The source sentence is translated into a linear, lexical constituency tree of the target sentence which is further used to generate the output sentence. Eriguchi et al. (2019) used the source-side phrase structures and introduced a tree to string approach in NMT.

We introduce a different approach where we do not include any additional monolingual data to train the NMT model, as was done in the existing literature, mentioned above. Rather, we extract the syntactic phrases from the target-side sentences of the original parallel data and use it along with the original corpus for training the NMT system. Instead of learning from the additional monolingual target side samples, our model learns from the phrases obtained from the original data itself. It provides better evidence for the NMT model to learn the parameters.

3 Motivation

Our work is motivated by the idea of improving the translation performance of the low-resource language pairs.

A low-resource scenario is not only the absence of a significant amount of parallel data for a language pair only, but for a specific domain too. For example, Indian languages like Hindi, Malayalam, Telugu etc. in comparison to English and European languages, may be considered as low-resource. The situation is even worse for some specialized domains like judicial and health, for which obtaining parallel corpus and even the monolingual corpus is a challenge. To deal with this situation, we propose a data augmentation strategy that makes use of syntactic phrases as additional examples for training.

The rationales behind are the following:

1. Syntactic phrases extracted from the constituency tree are linguistically more sound compared to the phrases extracted from the phrase table.
2. The technique does not require any additional monolingual data. We have empirically observed that it further improves the translation quality, even if used with additional back-translated synthetic data.
3. Extracting syntactic phrases from the available target sentences of authentic data provides in-domain data which are similar to the style of authentic corpus only.
4. In a domain like judicial, sentences are long (Iyer 2020), and augmenting multiple linguistically correct syntactic phrases of different sizes of a target sentence would help model to learn that sample efficiently, and they essentially diversify the training data.
5. English language (word order Subject–Verb–Object, i.e. SVO) is syntactically diverse in comparison to Hindi, Telugu and Malayalam languages (word order, SOV). To deal with the syntactic divergence, Collins et al. (2005) used reordering at the source side, Ramanathan et al. (2011) reordered the source sentence on clause basis and Zhou et al. (2019) used reordering at the target side to decrease the syntactic distance while translating. These techniques help the model to learn

the source-target pair in a better way. As we are sub-sampling a target sentence into various syntactically correct phrases with different lengths, we try to provide the learning model more (reasoning) knowledge about that sentence-pair.

6. Translating a small text sequence is comparatively easier than translating a large text sequence. This is the reason why our proposed idea of back-translating syntactic phrases is more efficient in terms of data augmentation where better back translation systems are not available as such. In these cases, our proposed idea can be seen as an efficient alternative to the idea of incorporating back-translated monolingual data.

4 Augmenting training data with syntactic phrasal-segments

This section presents our proposed generalized data augmentation method. As mentioned above in Sect. 1, in this study we consider three low-resource language-pairs whose target-side is English. To the best of our knowledge, there is no freely available constituency parser for any of the source-side languages (i.e. Hindi, Malayalam, and Telugu) of our chosen translation-pairs. Hence, in order to obtain linguistically coherent phrases, we rely on the target-side language English. First, we parse the English sentences of the bilingual authentic training corpus using Stanford parser[1] and obtain the corresponding constituency trees.

As for extraction, we considered three primary phrase structure types: noun phrase (NP), verb phrase (VP) and preposition phrase (PP). We consider all such phrases irrespective of their sizes. As an example, we see from Fig. 1 that terminals 'the', 'Imperial', and 'Library' under an NP node form a noun phrase "the Imperial Library". We extracted all potential noun, verb and prepositional phrases from the English sentences. In order to obtain source counterparts of the extracted English phrases, we back-translate them with a target-to-source language NMT system (c.f. Sect. 8). Note that the target-to-source NMT system was built on the original source–target parallel corpus. This results in a phrase-level synthetic corpus which is then added to the original parallel corpus. Finally, the source-to-target NMT system is trained on the augmented corpus.

In order to illustrate our synthetic corpus creation process via an example, we took the following source-target sentence-pair from Hindi-to-English training corpus: " इस अधिनियम को इंपीरियल लाइब्रेरी का नाम राष्ट्रीय पुस्तकालय में बदलने के लिए पारित किया गया था।" (*is adhiniyam ko impeeriyal laibreree ka naam raashtreey pustakaalay mein badalane ke lie paarit kiya*) and 'This act was passed to change the name of the Imperial Library to National Library.'

The parse tree of the English sentence is shown in Fig. 1, and the phrases extracted from the tree are shown in the first column of Table 1. The translations produced by our English-to-Hindi back-translation NMT system are shown in Table 1.

We observe that translations of the English phrases produced by our target-to-source NMT models were good (syntactically and semantically) in quality in all the

[1] https://nlp.stanford.edu/software/lex-parser.shtml.

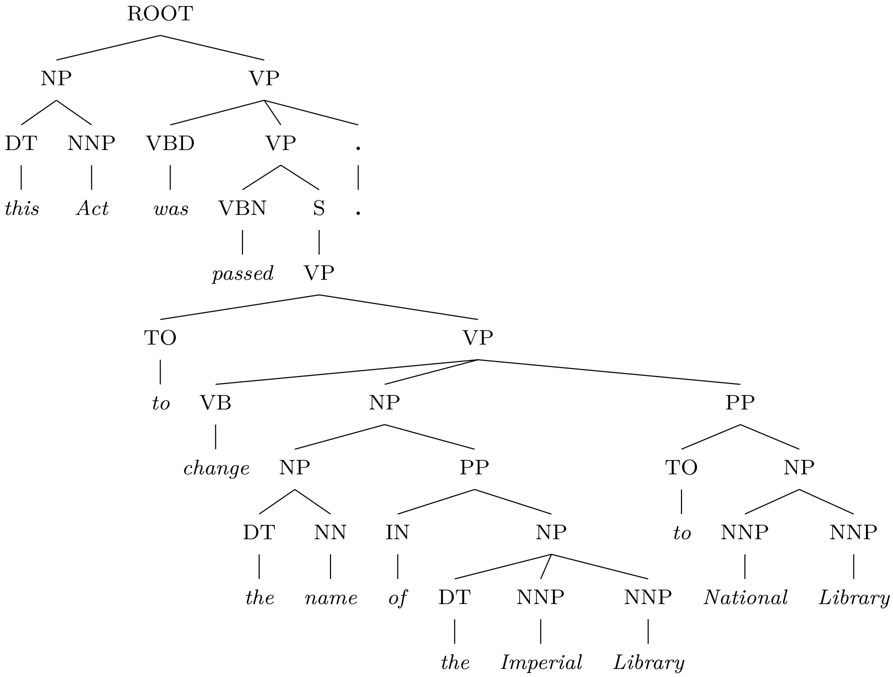


Fig. 1 Parsed tree

three translation tasks. Note that we extract the English phrases from the source-side of the training corpora on which our back-translation MT systems were built. This could be the reason why the translations of the English phrases were excellent in quality. We also manually analyze 100–150 output samples by randomly choosing them. We find that the quality is good in terms of syntax and semantics. We also adopt an alternative method in order to obtain the source counterparts of the extracted English phrases, which were found to be less effective. We discuss this matter in Sect. 9.1.3.

Currey et al. (2017) generated synthetic parallel sentences by copying target sentences to the source. Following (Currey et al. 2017), we created a synthetic corpus by copying the extracted English phrases to the source in each translation task. In this way, we have a parallel corpus where source and target samples are identical. We call this synthetic corpus *copied corpus* (CC). We carried out experiments by adding the copied corpus to the original authentic corpus. Both synthetic corpus creation strategies are shown in Fig. 2. We also conducted experiments by adding both synthetic and copied corpus together to the authentic training corpus.

Table 1 English phrases from the parse tree with their Hindi translation

English phrase	Hindi translation
National Library [NP]	राष्ट्रीय पुस्तकालय (national library)
To National Library [PP]	राष्ट्रीय पुस्तकालय को (national library to)
The Imperial Library [NP]	इंपीरियल पुस्तकालय (imperial library)
Of the Imperial Library [PP]	इंपीरियल पुस्तकालय के (imperial library of)
The name [NP]	नाम (name)
The name of the Imperial Library [NP]	इंपीरियल पुस्तकालय का नाम (imperial library of name)
Change the name of the Imperial Library to National Library [VP]	के नाम परिवर्तन करने के लिए राष्ट्रीय (of name change to national पुस्तकालय इंपीरियल पुस्तकालय library imperial library)
To change the name of the Imperial Library to National Library [VP]	के नाम में परिवर्तन करने के लिए राष्ट्रीय (of name in change to national पुस्तकालय इंपीरियल पुस्तकालय library imperial library)
Passed to change the name of the Imperial Library to National Library [VP]	के नाम में परिवर्तन करने के लिए पारित इंपीरियल (of name in change to passed imperial पुस्तकालय के लिए राष्ट्रीय पुस्तकालय library to national library)
Was passed to change the name of the Imperial Library to National Library [VP]	के लिए किया गया था इंपीरियल पुस्तकालय (to imperial library का नाम को राष्ट्रीय पुस्तकालय of name to national library)
This Act [NP]	यह अधिनियम (this act)

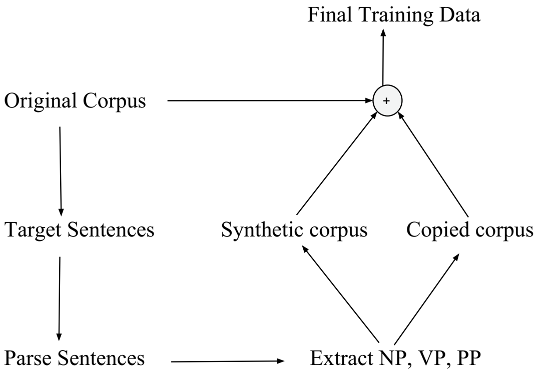


Fig. 2 Overall architecture. '+' means append

5 Datasets

This section provides the details of the datasets that we use for our experiments. In Table 2, we show the dataset statistics. The top rows represent the statistics of the data sets for the Hindi-to-English MT task. In fact, we have three MT subtasks for the Hindi-to-English translation. For the first subtask, we use the WMT14 Hindi–English data set (Bojar et al. 2014b). For the second and third sub-tasks we

Table 2 Corpus statistics

	Sentences	Words (SL)	Words (TL)
Hindi–English WMT14			
Train set	263,654	3,330,273	3,033,689
Train set (vocab)		104,017	112,345
Newsdev2014	520	10,181	10,317
Newstest2014	2507	63,904	55,818
Hindi–English Judicial			
Train set	5000	129,971	121,430
Train set (vocab)		9327	8360
Development set	1000	15,654	16,342
Test set	1000	15,776	16,876
Hindi–English Health			
Train set	23,000	418,853	391,943
Train set (vocab)		19,999	17,257
Development set	1000	16,123	17,976
Test set	1000	16,897	18,065
Malayalam–English			
Train set	359,268	2,229,593	3,048,936
Train set (vocab)		252,530	58,224
Development set	500	8,560	10,782
Test set	1,000	15,981	20,160
Telugu–English			
Train	22,165	126,310	153,725
Train set (vocab)		9930	12,729
Development set	500	8154	9495
Test set	1000	15,322	18,053

use two domain-specific data sets (health (Jha 2010) and judicial (Kunchukuttan et al. 2018)). In Table 2, we report the number of training, development and test set sentences for each of the sub-tasks. For the WMT14 translation task we use the standard development and test sets from WMT14 (Bojar et al. 2014a). As far as the other two MT tasks (health and judicial) are concerned, the samples for development and test sets were taken randomly from the respective data sets.

We use the WAT18 data sets[2] (Nakazawa et al. 2018) for the Malayalam-to-English and Telugu-to-English translation tasks. For each task, we randomly pick 500 and 1,000 sentences for the development and test sets, respectively. The bottom half of Table 2 represents the corpus statistics for the Malayalam-to-English and Telugu-to-English MT tasks.

[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/indic_languages_corpus.tar.gz.

6 Baseline neural machine translation models

6.1 Attention based recurrent neural network (ARNN)

The goal of encoder-decoder based recurrent translation models is to translate a sequence of input tokens into a sequence of output tokens using a deep neural network. There are two key components in this model: encoder and decoder. The encoder encodes the source sequence into a hidden representation and the decoder uses this representation to generate output tokens from left to right one at a time. The encoder-decoder model has a limitation that it fails to capture the dependency if the input sentence length increases. The encoder of this model tries to encode all the tokens information into a single hidden representation, which is not efficient, specifically in case of the longer sequences. To overcome this limitation, Bahdanau et al. (2015) introduces the idea of attention which focuses on the whole input sequence while generating the tokens at the target side. This mechanism is called "attention based recurrent neural network" (ARNN) which gives the weights to the input tokens as per their contribution in generating a target token.

For a given input and output sequence pair [x, y], where $x = (x_1, x_2, \dots, x_m)$ is a sequence of input tokens and $y = (y_1, y_2, \dots, y_n)$ is a sequence of output tokens. The probability of the ith output word y_i is calculated as in Eq. (1):

$$p(y_i|y_1, \dots, y_{i-1}, x) = f(y_{i-1}, s_i, c_i) \tag{1}$$

Here, s_i and c_i are the ith hidden state at the decoder side and context vector, respectively. As shown in Eq. (1), in NMT, at the time of decoding, the next predicted output token y_i depends on the previous output tokens $y_1, \dots, y_{i-1}$.

6.2 Google's transformer model (GT)

In recurrent neural networks (ARNN), the hidden representation at any timestamp depends on the previous hidden state. Vaswani et al. (2017) proposed the transformer network which removes the recurrent networks as found in the ARNN model and depends on self-attention and allows for parallel computation of hidden states at all timestamps in the encoder and the decoder. However, in the absence of the recurrent network, to handle the position of the tokens in the input sequence, a new vector named as positional encoding is appended with each input embedding before the encoder calculates the hidden representations. The encoder is composed of some identical layers which are composed of two sub-layers: *(i)* a multi-head self-attention layer and *(ii)* a position-wise feed-forward network layer. Similar to the encoder, the decoder is composed of several identical layers. In addition to the two sub-layers in the encoder, the decoder adds a third sub-layer, which calculates multi-head attention over the encoder's output. These sub-layers are followed by a normalization layer. The decoder in GT performs similarly to the decoder in ARNN and generates one token at each timestamp with the help of a softmax layer.

6.3 Convolutional sequence to sequence neural network (CSSNN)

convolutional Neural Network (CNN) architecture for seq2seq learning (Gehring et al. 2017) applies Gated Linear Units (GRU) (Dauphin et al. 2017) and residual connections. It also applies separate attentions to each decoder layer and demonstrates that each attention adds very little overhead. In this architecture, encoder and decoder RNNs are fully replaced by CNN. Here, the computations over all the elements can also be fully parallelized during training.

7 Experimental setup

As already mentioned, we make use of three state-of-the-art NMT paradigms, ARNN (Bahdanau et al. 2015), GT (Vaswani et al. 2017) and CSSNN (Gehring et al. 2017). Additionally, in order to compare the neural approaches with previous state-of-the-arts, we also build a PB-SMT system (Koehn et al. 2003). As far as the back-translation of the extracted English phrases is concerned, we build two target-to-source MT systems, a PB-SMT system and a GT system for each task. Since there are three MT sub-tasks for the Hindi-to-English translation, we had to build all the MT system types for each sub-task.

We tokenize and true-case the sentences and remove those sentences whose number of words is more than 80. In order to tokenize Hindi, Malayalam and Telugu sentences, we use Indic_NLP_library[3]. Note that Hindi, Malayalam and Telugu are unicase languages.

For PB-SMT training, we use the Moses toolkit (Koehn et al. 2007). Our PB-SMT model includes: (i). forward and backward lexical and phrase probabilities, (ii). 8 lexicalized reordering probabilities, (iii). 5-gram language model (Kneser–Ney smoothing (Kneser and Ney 1995)), (iv). 5 OSM features (Durrani et al. 2011), and (v). distortion penalties and word-count. In our experiments, we use GIZA++ toolkit[4] (Och and Ney 2003) to train the word alignment models, *grow-diag-final-and* algorithm of (Koehn et al. 2003) is used for phrases extraction, Kneser–Ney smoothing is used for phrase scoring. For decoding, the distortion limit is 12.

To develop our NMT models, we use the Sockeye toolkit[5] (Hieber et al. 2018). The tokens in the train, test and development sets are segmented into subword units using the byte pair encoding (BOE) technique (Sennrich et al. 2016c). Since English is written in Roman alphabet, and Hindi, Malayalam and Telugu are written in their respective scripts and there are no overlapping characters for languages, BPE is applied independently on the source and target side. We perform 32,000 join operations.

For the GT Vaswani et al. (2017), ARNN Bahdanau et al. (2015) and CSSNN Gehring et al. (2017), we keep the hidden layer size and word embedding dimension

[3] https://github.com/anoopkunchukuttan/indic_nlp_library.t

[4] http://www.statmt.org/moses/giza/GIZA++.html.

[5] https://github.com/awslabs/sockeye.

Table 3 Performance of target-to-source PB-SMT and NMT systems (GT) on BLEU

EN-to-HI			EN-to-ML	EN-to-TE
WMT14	Judicial	Health		
PB-SMT				
8.93	**22.21**	**19.93**	18.32	**23.65**
GT				
8.56	20.24	18.24	**19.92**	14.21

Bold values indicate the best scores among all

Table 4 Numbers of train set sentences and extracted English phrases

Dataset	Sentences	Phrases
WMT14	263,654	595,969
Judicial	5000	81,308
Health	23,000	91,189
ML-EN	359,268	557,138
TE-EN	22,165	53,702

as 512. Max sentence length is 80, batch size is 4000 tokens, and learning rate is 0.0002. To stop the training, we use the early-stopping criteria based on the development set. We use the Adam Kingma and Ba (2015) optimizer. In CSSNN, dropout over the inputs of the convolutional blocks is used. In GT, the number of encoder-decoder layers are 6-6 and the number of the attention heads is 8.

8 Target-to-source MT systems for back-translation

As pointed out above in Sect. 4, we translate the extracted English phrases using the target-to-source MT system. For this, we build a PB-SMT (Koehn et al. 2003) system and a GT (Vaswani et al. 2017) system for each of the translation tasks. In Table 3, we present the comparative performance of the PB-SMT and GT system in terms of BLEU. We see from Table 3 that for all our translation tasks we found that our PB-SMT systems statistically significantly outperform our NMT systems barring the English-to-Malayalam task where we see GT significantly outperforms the PB-SMT. This is the reason why we chose PB-SMT over GT for back-translation except the English-to-Malayalam task for which we chose GT for back-translation. Note that we use bootstrap resampling methods (Koehn 2004) to perform statistical significance tests. If the improvement in the system's performance is at a confidence level above 95% then the improvement is considered to be statistically significant.

Translating from the morphologically-poor (e.g. English) to morphologically-rich language (e.g. Hindi, Malayalam, and Telugu) is arguably more challenging than the other way round. This is also the case with us as we see from Table 3 that the BLEU scores of the MT systems (English-to-Hindi, -Malayalam and -Telugu PB-SMT and

GT systems) are worse than those of the Hindi, Malayalam and Telugu-to-English MT systems which are reported in the results section (c.f. Tables 5 and 11; Sect. 9).

In Table 4, we show the number of training set sentences from which the English phrases were extracted. The last column of the table shows the total number of extracted English phrases. The English phrases were back-translated using the respective PB-SMT system.

9 Results and discussions

In this section we report our experimental results and discussions on the Hindi-to-English MT task (in Sect. 9.1) and Malayalam-to-English and Telugu-to-English Translation (in Sect. 9.2). We also give analysis on MT translation in Sect. 9.3.

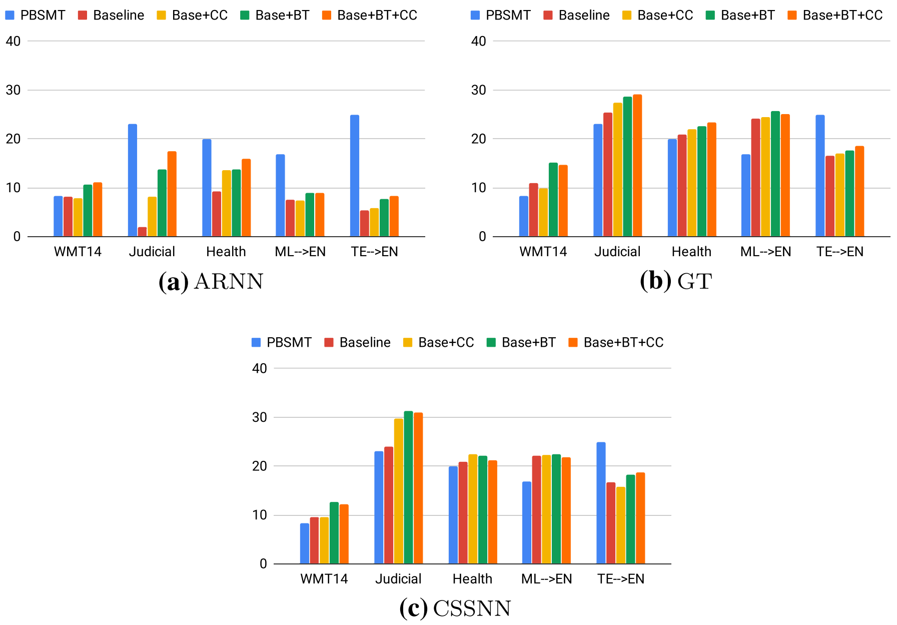
9.1 Hindi-to-English translation

9.1.1 Linguistically motivated data augmentation

We evaluate our proposed approach on three different kinds of datasets, *viz.* WMT14, Judicial and Health. We report the evaluation results for each of the tasks in Table 5, which are grouped by the MT system types: ARNN, CSSNN, and GT. The third column (Base) of the table represents the BLEU scores of our three vanilla NMT baseline systems (see baseline experimental setups in Sect. 7). Additionally, we build a vanilla PB-SMT baseline system for each of the MT tasks. The intuition underpinning this is that we intend to compare PB-SMT and NMT and see how they perform in the low-resource translation settings. We report the performance of the PB-SMT systems in terms of BLEU in the last rows of Table 5. When we compare the numbers of the third column (Base), we see that GT is the clear winner among the four baselines in the WMT14 and Judicial MT tasks. We also see from that column that the GT and CSSNN baseline systems can be regarded as comparable in the Health MT task as we obtained almost the same BLEU scores on the test set for them.

As mentioned in Sect. 4, we applied two different strategies in order to append phrase-level synthetic corpus to the training data: (i) back-translating the extracted English phrases to the source with a target-to-source MT system; from now, we call this approach BT and (ii) copying the extracted English phrases verbatim to the source; from now, we call this approach CC. The fourth (Base + BT) and fifth (Base + CC) columns of Table 5 represent those NMT systems, built on the training data, which were prepared following the CC and BT strategies, respectively. We also applied the CC and BT strategies together to the baseline setup. The results of the combined setup are presented in the last column (Base + BT + CC) of Table 5.

As far as the WMT14 translation task is concerned, we see from Table 5 that the CC strategy does not bring any improvement. However, for the BT strategy all the NMT models (i.e. Base + BT setup) produced statistically significant improvements in terms of BLEU over the baselines. When we applied CC and BT collaboratively



(a) ARNN

(b) GT

(c) CSSNN

Fig. 3 BLEU scores of the different ARNN, GT and CSSNN systems in the Hindi-to-English translation

Table 5 Evaluation results for Hindi-to-English translation

System	Base	Base + CC	Base + BT	Base + BT + CC
ARNN				
WMT14	8.1	7.89	10.58	**11.05**
Judicial	1.87	8.15	13.81	**17.50**
Health	9.28	13.54	13.79	**15.95**
CSSNN				
WMT14	9.47	9.50	**12.59**	12.13
Judicial	23.98	29.72	**31.22**	30.97
Health	20.93	22.43	22.06	21.27
GT				
WMT14	**10.95**	9.92	**15.17**	14.72
Judicial	**25.36**	27.47	28.68	**29.13**
Health	**20.95**	22.02	22.52	**23.33**
PB-SMT				
WMT14	8.24		10.35	
Judicial	23.13		24.26	
Health	20.02		21.34	

Bold values indicate the best scores among all

Table 6 Comparing performance of models over newstst2014 (Hindi-English)

	Base	Base + BT (proposed)	Fadaee et al. (2017)	Zhu et al. (2019)
WMT14	10.95	15.17	13.05	14.19

(i.e. Base + BT + CC setup), we see that the combined approach does not help. In sum up, we obtain the best BLEU points with GT (an absolute of 4.22 points corresponding to 2.78% relative improvement in terms of BLEU over a GT baseline) when we apply the BT strategy.

When we compare the BLEU scores in the Judicial and Health MT tasks, we see that both the strategies (CC and BT), individually and collaboratively, are effective in improving the MT system's translation quality. However, in the Judicial translation task, we obtain the best BLEU score on the test set (an absolute of 7.24 points corresponding to 30.2% relative improvement in terms of BLEU over the baseline) with CSSNN when the BT strategy is applied to the baseline setup. For the Health MT task, GT becomes the winner when both BT and CC strategies are applied together, and the best setup (Base + BT + CC) produces a 23.33 BLEU points on the test set, which corresponds to an absolute 2.38 points gain (a 10.2% relative improvement) over a GT baseline. In Fig. 3, we show the BLEU scores of different MT systems for the three system types: ARNN, CSSNN, and GT.

We also perform experiments with syntactic phrase augmented data (Base + BT) using PB-SMT. As shown in Table 5, we obtain 2.11, 1.13 and 1.72 BLEU points improvement over the PB-SMT baseline using WMT14, judicial and health dataset, respectively. We also compare the performance of the proposed method with the word augmentation approaches Fadaee et al. (2017), Zhu et al. (2019). In Table 6, we can see that augmenting syntactic phrases demonstrates better performance compared to the data augmentation methods over the WMT14 Hindi-English dataset. Evaluation shows that our model yields a significant improvement of 2.12 and 0.98 BLEU points over the models Fadaee et al. (2017) and Zhu et al. (2019), respectively.

For each data domain (WMT14, judicial and health), we perform significance test[6] to verify whether BLEU gains for the MT systems (ARNN, CSSNN, and GT) with Base + BT and Base + BT + CC setups over the respective baselines are statistically significant. We find that the results are significant with 95% CI (which is< 0.05).

9.1.2 Human evaluation

Along with automatic evaluation, we also perform human evaluation to study the translation quality from the user's point of view. Sentences are assigned with

[6] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/analysis/bootstrap-hypothesis-difference-significance.pl.

Table 7 Average adequacy and fluency score

	Adequacy	Fluency
Base (WMT14)	2.13	2.47
Base + BT (WMT14)	**2.65** (+ 0.52)	**2.75** (+ 0.28)
Base (Judicial)	**2.48**	**2.79**
Base+BT (Judicial)	**3.20** (+ 0.72)	**3.41** (+ 0.62)

Bold values indicate the best scores among all

Table 8 Evaluation results using *n*-gram phrases on WMT14 MT task

	ARNN	CSSNN	GT
Base	8.10	9.47	10.95
Base + GIZAphrase	8.51	9.63	**11.83**

Bold values indicate the best scores among all

adequacy and fluency scores. We use the following ratings for *'Adequacy'* and *'Fluency'*:

0: *Incorrect*, **1:** *Almost Incorrect*, **2:** *Moderately Incorrect*, **3:** *Almost Correct*, **4:** *Correct*.

We select 100 random samples from the test sets (WMT14 and judicial domain) and employ two experts to assign the adequacy and fluency scores for each test sample. The language experts are post-graduates in linguistics and have good knowledge in the language. Table 7 shows the average rating for the 'Base' and 'Base + BT' models.

9.1.3 Experiments with *n*-gram phrases

We pointed out in Sect. 4 that in order to obtain the source counterparts of the extracted English phrases, we adopted an alternative method to the back translation strategy. Since we extract the English phrases from the target-side of the bilingual training corpus, we easily obtain source–target word alignment information using GIZA++. Then, we extract bilingual *n*-gram phrases using *grow-diag-and-final* algorithm (Koehn et al. 2003), and consider those *n*-gram phrase-pairs whose target parts are syntactic English phrases. For extraction we set maximum phrase length (i.e. *n*) to the number of words of the longest English syntactic phrase. As an example, we extracted 7,002,108 *n*-gram phrase-pairs from the WMT14 training corpus following the phrase-extraction algorithm of Koehn et al. (2003). We found that the target sides of 210,366 phrase-pairs (out of 7,002,108 phrase-pairs) were found in the list of the extracted syntactic English phrases (595,969 entries). This number (i.e. 210,366 entries) is approximately 3% of the total number of *n*-gram phrase-pairs. This is quite reasonable since the *grow-diag-and-final* algorithm extracts non-linguistic *n*-gram phrase-pairs that are *consistent* with the GIZA++ word alignments (Koehn et al. 2003).

We carried out a set of experiments after adding the extracted *n*-gram phrase-pairs with the training corpus. We call this experimental setup as Base + GIZAphrase.

Table 9 Evaluation Results from the back translation experiments

	Base	1:1	1:2	1:4
GT	10.95	12.94	14.13	14.66
CSSNN	9.47	10.03	10.53	11.34
ARNN	8.1	8.92	9.79	10.22

This investigation is carried out on the WMT14 data set. We report the test set BLEU scores in Table 8. As can be seen from Table 8, addition of the *n*-gram phrases to the training data brings improvements on the test set accuracy. The improvements over the baselines are not statistically significant for the ARNN and CSSNN systems. Although we obtain statistically significant BLEU gain (an absolute 0.88 BLEU points corresponding to 8.03% relative improvement) on the test set for GT, the gain is considerably lower than that we obtained with the Base + BT setup (cf. Table 5).

9.1.4 Investigating back translation

Sennrich et al. (2016b) presented a simple method for exploiting target-side monolingual data in NMT. They created a synthetic parallel data from the target-side monolingual corpus via back-translation. Although this is a very simple method, it provides an efficient way to use target-side monolingual data in NMT training and is regarded as a benchmark approach to reduce the data sparsity, especially in resource-poor scenarios. Fadaee and Monz (2018) observed that the NMT model trained on training data composed of 1:4 authentic-to-synthetic ratio achieved slight improvements over the model that was trained on the training data of 1:1 authentic-to-synthetic ratio.

We carried out a set of experiments by adding synthetic data to the training corpus. Following Fadaee and Monz (2018), we explored different ratios for combining authentic and synthetic data to test which ratio achieves the best performance. For creating synthetic data, we use monolingual data from Europarl corpus (Koehn 2005). We obtain the BLEU scores to evaluate the different NMT systems that were trained on the training data of different authentic-to-synthetic ratio, and mention them in Table 9. As for the WMT14 MT task, 1:4 is found to be the most effective ratio for combining authentic and synthetic data.

9.1.5 Linguistically motivated data augmentation in collaboration with back translation

In Section 9.1, we presented the evaluation results obtained by applying our proposed linguistically motivated data augmentation approach to NMT. Base + BT is found to be the best setup as per the BLEU scores shown in Table 5 as far as the WMT14 MT task is concerned. From our back-translation experiments we found that 1:4 is the most effective ratio for combining authentic and synthetic data. These back translation experiments were also conducted on the WMT14 data sets. Further, we tested our proposed data augmentation approach in collaboration with the

Table 10 Evaluation Results obtained applying our approach in collaboration with the back translation

	Base	Base + BT	Base + BackT(1:4)	Base + BackT(1:4) + BT
GT	10.95	15.17	14.66	17.31
CSSNN	9.47	12.59	11.34	14.16
ARNN	8.1	10.58	11.22	11.89

Table 11 Evaluation results for Malayalam (Ml)-to-English (En) and Telugu (Te)-to-English (En) translation

System	Base	Base + CC	Base + BT	Base + BT + CC
Ml-to-En				
GT	24.10	24.53	25.68	25.07
CSSNN	22.19	22.31	22.36	21.85
ARNN	7.49	7.32	8.99	8.89
PB-SMT	16.81		18.31	
Te-to-En				
GT	16.46	16.93	17.62	18.63
CSSNN	16.67	15.69	18.22	18.74
ARNN	5.32	5.76	7.66	8.33
PB-SMT	24.93		26.05	

back-translation approach (Sennrich et al. 2016a). In Table 10, we report the BLEU scores to evaluate the NMT systems on the WMT14 test set for the collaborative setup. We see from Table 10 that the collaborative setup turns out to be the best-performing in our low-resource experimental settings, and both strategies were found to be complementary to each other.

9.2 Malayalam-to-English and Telugu-to-English translation

In addition to the Hindi-to-English translation, we carried out experiments on the Malayalam-to-English and Telugu-to-English translation. This section shows the results from the Malayalam-to-English and Telugu-to-English translation tasks.

9.2.1 Linguistically motivated data augmentation

As in above, we carried out experiments considering various setups (Base, Base + CC, Base + BT and Base + BT + CC) in the Malayalam-to-English task. We obtain the BLEU scores to evaluate the different MT systems on the test set, which are shown in Table 11. In the Ml-to-En MT task, Base + BT is found to be the best-performing setup. With this, we achieve the absolutes of 1.58 points (corresponding to 6.55% relative), 0.17 points (corresponding to 0.76% relative) and 1.50 points (corresponding to 20.02% relative) improvements in terms of BLEU points on the test set over the GT, CSSNN and ARNN baselines, respectively. The improvements over the GT and ARNN baselines are statistically significant. As far as Base + CC setup is concerned, we found that none of the improvements is statistically significant.

In this task, we also see that the Base + BT + CC does not help much. Further in the Telugu-to-English task, Base + BT + CC turns out to be the best setup for this task. It brought absolute gains by 2.17 (corresponding to 13.18% relative), 2.07 (corresponding to 12.41% relative) and 3.01 (corresponding to 56.57% relative) points on the test set over the GT, CSSNN and ARNN baselines, respectively. We found that the gains are statistically significant. The NMT systems which were built as per the Base + BT setup also surpassed the respective baselines with substantial margins in terms of BLEU. We found that these gains are statistically significant with 95% confidence level (which is< 0.05). We refer the reader to Fig. 3, where we show the BLEU scores of different MT systems for all MT types, ARNN, CSSNN, and GT. In Table 11, we can see that the BLEU score for the PB-SMT system is mentioned in the last row.

As in above, we built a PB-SMT baseline system for Ml-to-En and Te-to-En MT tasks. We can see that for Ml-to-En task, PB-SMT does not work well in comparison to the best baseline neural models, GT and CSSNN while as opposed to Hi-to-En and Ml-to-En MT tasks for Te-to-En task, PB-SMT significantly outperforms the NMT baseline models with the margins of more that 8 BLEU points. As shown in the Table 13, for the Te-to-En task, even the test set BLEU score of our best neural model is not comparable to that of the PB-SMT system. However, since in this work our primary objective is to investigate the training data augmentation methods (ours and the existing benchmarks) with the state-of-the-art neural models in low-resource and extremely low-resource conditions, we test them in the Telugu-to-English task too despite the fact that PB-SMT outperforms NMT in this case. In particular, for Ml-to-En tasks, the GT baseline produces a 7.29 BLEU gain over the PB-SMT baseline, which is statistically significant with 95% confidence level (which is< 0.05). We also perform experiments with syntactic phrase augmented data (Base + BT) using PB-SMT. As shown in Table 11 we obtain 1.50 BLEU point improvement over the PB-SMT baseline.

9.2.2 Experiments with *n*-gram phrases

As in Hindi-to-English MT tasks, we extract *n*-gram phrase-pairs from the training corpus in the Malayalam-to-English and Telugu-to-English MT tasks. The number of phrase-pairs that were taken from the corpus are 8,241,074 and 717,946 respectively. We found that for Malayalam-to-English, the target-sides (English) of 376,108 phrase-pairs (out of 8,241,074 *n*-gram phrase pairs) were found in the list of extracted syntactic English phrases (557,138). This is approximately 4.6% of the total number of *n*-gram phrase-pairs. Also for Telugu-to-English, the target-side (English) phrases of 9,381 phrase-pairs (out of total 717,946 *n*-gram phrase-pairs) were found in the list of extracted syntactic English phrases (53,702). This is approximately 1.3% of the total number of *n*-gram phrase-pairs. We carried out experiments with adding *n*-gram phrase-pairs extracted using GIZA++ alignments

Table 12 Evaluation results using *n*-gram phrases on Malayalam (Ml)-to-English (En) MT and Telugu (Te)-to-English (En) tasks

	GT	CSSNN	ARNN
Ml-to-En			
Base	24.10	22.19	7.49
Base + GIZAphrase	24.76	22.28	8.02
Te-to-En			
Base	16.46	16.67	5.32
Base + GIZAphrase	16.33	16.69	5.46

Table 13 Evaluation results from the back translation experiments

	Base	1:1	1:2	1:4	Base	Base + BT +	Base + BT + CC
					+ BT	BackT(1:4)	+ BackT(1:4)
Ml-to-En							
GT	24.10	23.17	24.03	24.34	**25.68**	**26.19**	–
CSSNN	22.19	21.27	21.69	22.06	22.36	**23.87**	–
ARNN	7.49	7.78	8.54	9.02	**8.99**	**11.34**	–
Te-to-En							
GT	16.46	15.83	16.67	**17.91**	**17.62**	**19.67**	**20.24**
CSSNN	16.67	16.52	17.19	**18.08**	**18.22**	**19.63**	**20.07**
ARNN	5.32	4.31	5.64	**7.60**	**7.66**	**8.27**	**8.76**

Bold values indicate the best scores among all

to the authentic corpus. We obtain the BLEU scores to evaluate the MT systems on the test set, which are reported in Table 12. For Ml-to-En, the setup (Base + GIZAphrase) brought the gains of 0.66, 0.09 and 0.53 BLEU points on the test set over the GT, CSSNN and ARNN baselines, respectively and for Te-to-En, the setup (Base + GIZAphrase) brought the gains of 0.14, 0.02 BLEU points on the test set over the GT and CSSNN baselines respectively. We found that the improvement over the GT baseline for Ml-to-En is statistically significant.

9.2.3 Investigating back translation

We conduct a set of experiments by appending synthetic corpus to the authentic corpus on the Hindi-to-English translation (cf. Sect. 9.1.4). We also performed a similar set of experiments for the Malayalam-to-English and Telugu-to-English tasks. As a result, a set of MT systems was built on the training corpus compiled on the basis of various authentic-to-synthetic ratios. We obtain the BLEU scores to evaluate the MT systems on the test set, which are reported in Table 13. We can see from the table that

1. 1:1 and 1:2 authentic-to-synthetic ratios are not beneficial since we do not get consistent improvements for these setups.

2. For 1:4 authentic-to-synthetic ratio, we obtain the BLEU scores which are nearly comparable to those we obtained on the baselines.
3. As in above, for Malayalam-to-English translation, we also carry out experiments with a combined setup which is Base + BackT(1:4) + BT. With this, we achieve absolutes of 2.09 (corresponding to 8.67% relative), 1.68 (corresponding to 7.57% relative) and 3.85 (corresponding to 51.40% relative) points improvement on the test set over the GT, ASSNN, and ARNN baselines, respectively.
4. For Telugu-to-English translation, while comparing the numbers of sixth (Base + BT) and seventh (Base + BackT(1:4) + BT) columns of Table 13, we found that the NMT systems (GT, CSSNN, and ARNN) that represent Base + BaseT (1:4) + BT significantly outperform the respective NMT systems that represent Base + BT.
5. Further, for Telugu-to-English translation, we can see From Table 11 that the best test set BLEU scores are yielded with Base + BT + CC for this translation task. This led us to carry out experiments with an additional setup for this task: Base + BT + CC + backT(1:4). The test set BLEU scores for the NMT systems are shown in the final column of Table 13. We see from Table 13 that the Base + BT + CC + backT(1:4) setting further improves the performance and yields absolutes 3.78 (corresponding to 22.9% relative), 3.40 (corresponding to 20.4% relative) and 3.44 (corresponding to 64.6% relative) points improvement on the test set over the GT, CSSNN and ARNN baselines, respectively. We found these improvements are statistically significant.
6. We found that these improvements are statistically significant.

As in the Hindi-to-English task, we clearly see the BT and BackT setups are complementary to each other.

As in above, we also carry out experiments with a combined setup which is Base + BackT(1:4) + BT. With this, we achieve absolutes of 2.09 points (corresponding to 8.67% relative), 1.68 points (corresponding to 7.57% relative) and 3.85 points (corresponding to 51.40% relative) improvements in terms of BLEU on the test set over the GT, ASSNN, and ARNN baselines. We found that these improvements are statistically significant.

9.3 Analysis of the translation

This section presents the analysis on the translations generated by the different MT systems. For this analysis the source Hindi sentences were taken from the test sets of judicial and health domains. In Table 14, we show the source Hindi sentences and their translations produced by the different MT systems. For the first example, we took the Hindi source sentence from the judicial domain test set. Note that we consider PB-SMT and GT for this analysis. We see from the table that the English translations produced by both PB-SMT and NMT (GT) baselines are quite poor. We found the following problems with the translation produced by the GT baseline: (i). repetition of

Table 14 Example translations produced by different MT systems

	Output #1 (Judicial domain)
Source	पक्षकारों के वकील को सुन लेने पर और इस मामले के तथ्यों को ध्यान में रखते हुए , मुझे आक्षेपति आदेश में कोई दुर्बलता नहीं मलि रही है और तदनुसार , याचकिा खारजि की जाती है
Transliteration	Pakshakaaron ke vakeel ko sun lene par aur is maamale ke tathyon ko dhyaan mein rakhate hue , mujhe aakshepit aadesh mein koee durbalata nahin mil rahee hai aur tadanusaar , yaachika khaarij kee jaatee
Reference	Having heard counsel for the parties and taking into consideration the facts of this case , I find no infirmity in the impugned order and accordingly , the petition is dismissed
PB-SMT	Having heard counsel for the parties facts of this case and in view of the I no infirmity in the the impugned order do not find petition and accordingly , is accordingly dismissed.
Base	It is submitted that the petitioner has been filed by the petitioner that the petitioner has been filed by the petitioner in the date of Rs. 2000 / -
Base + CC	In view of the counsel for the parties and this decision of the parties , there is no merit in the learned Single petition
Base + BT	The counsel for the learned counsel for the parties and I could not find any infirmity in the facts of the impugned order
Base + BT + CC	In view of the learned counsel for the parties and its finding of the matter , there is no infirmity in the impugned order and accordingly , petition is dismissed
	Output #2 (Health domain)
Source	बच्चों में खाँसी के साथ कान में संक्रमण होने का जोखमि रहता है।
Transliteration	Bachchon mein khaansee ke saath kaan mein sankraman hone ka jokhim rahata
Reference	There is danger of occurrence of ear infection with cough in children
PB-SMT	With cough in children danger of occurrence of infection in the ear
Base	There is a lot of pain in the body in the body in the body
Base + CC	In children there is danger of infection in the ear pain during cough
Base + BT	In children there is danger of infection in the ear with cough
Base + BT + CC	There is danger of infection in the ears with cough in children

phrases: *'the petitioner'*, dropping words: पक्षकारों (pakshakaaron), वकील (vakeel), मामले (maamle) and several incorrect lexicon selections. We also see that GT with *Base+CC* setup produces a fluent translation; however, there are some important words missing in the translation, e.g. 'याचकिा खारजि की जाती है' (yaachika khaarij k. With the Base + BT + CC setup GT produces a far better translation as we see it translates tokens with right ordering and the translation is fluent too.

We show another source sentence (Hindi) and its English translations in the bottom half of Table 14. This time, we pick the sentence from the health test set. As can be seen from the table, GT produces better translations with the *Base + BT* and Base + BT + CC setups than those by the baselines.

We also see from Table 14 that, in both the examples, the translations produced by the PB-SMT model are quite adequate, but less fluent mainly due to wrong word order.

We see a contrasting behaviour with GT. The translations produced by GT are more fluent, but lack adequacy.

10 Conclusion and future work

For training a NMT system a huge amount of training corpus is required. Such a corpus is not readily available for many language pairs and domains. In the proposed work, we have introduced an approach to improve the translation quality of an NMT system under the low-resource scenario by augmenting linguistically sound syntactic phrases generated by the parse trees of target side training sentences. We have extracted the noun phrases, verb phrases, and prepositional phrases, from the parse tree of target side sentences, and used the back-translation technique to generate parallel phrases for the source side. We have used this synthetic data of phrase pairs as additional training samples. We have shown empirically that our proposed technique of augmenting syntactic phrases to the original training data improves the baseline NMT system significantly for low resource language pairs like Hindi-to-English, Malayalam-to-English and Telugu-to-English. We also did qualitative analysis by performing human evaluation for Hindi–to–English translation (over WMT14 and judicial domain testset). Human evaluators calculated the average adequacy and fluency scores for the generated English outputs. It was found that our proposed model is more effective compared to the baseline.

As a continuation of this work, we have summarized the following research directions to work on in the future:

- We intend to include more low-resource languages from different language families in our study. We also aim to consider the narrow specialized domains in our investigation in order to observe how well our approach works in such scenarios.
- Currey et al. (2017) pointed out that the back-translation method (Sennrich et al. 2016b) is not very effective in low-resource scenarios where a good back-translation model is difficult to build. In future, we want to test our approach on some of those low-resource translation-pairs for which the back-translation engine is hard to build.
- We plan to extract syntactic phrases from both the source and target sides of the parallel training corpus and consider both types in our approach. It would be interesting to see their impacts on the translation quality.
- Given the fact that the constituency parsing is available for a handful of high resource languages (e.g. English, French), we aim to explore Hiero (Chiang 2005) in our work, which generates unsupervised hierarchical structures for the source and target sentences of a given parallel corpus.
- In this work, we compared our proposed approach with the benchmark strategies that are commonly used to minimise data sparsity in NMT (Sennrich et al. 2016b; Currey et al. 2017), and further, we explored the possibility of applying them together. In future, we aim to explore applying other existing benchmark

augmentation strategies individually or in collaboration with ours, e.g. self-training and back-translation in iterative fashion as in (Chen et al. 2019).

- We aim to explore more on the topics of morphological structures and character level encoding.

Acknowledgements Authors gratefully acknowledge the support of the project “Hindi to English Machine Aided Translation (HEMAT) for the Judicial Domain”, sponsored by TDIL, Meity, Govt. of India. Asif Ekbal gratefully acknowledges the Young Faculty Research Fellowship (YFRF), 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 MediaLab Asia).

References

Aharoni R, Goldberg Y (2017) Towards string-to-tree neural machine translation. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, pp 132–140

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: International conference on learning representation (ICLR)

Bojar O, Buck C, Federmann C, Haddow B, Koehn P, Leveling J, Monz C, Pecina P, Post M, Saint-Amand H, Soricut R, Specia L, Tamchyna A (2014a) Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the ninth workshop on statistical machine translation, Association for Computational Linguistics, Baltimore, pp 12–58

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

Chen P-J, Shen J, Le M, Chaudhary V, El-Kishky A, Wenzek G, Ott M, Ranzato M (2019) Facebook AI’s WAT19 Myanmar-English translation task submission. In: Proceedings of the 6th workshop on Asian translation, Hong Kong, pp 112–122

Chiang D (2005) A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, pp 263–270

Collins M, Koehn P, Kučerová I (2005) Clause restructuring for statistical machine translation. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, pp 531–540

Currey A, Miceli BAV, Heafield K (2017) Copied monolingual data improves low-resource neural machine translation. In: Proceedings of the second conference on machine translation. Association for Computational Linguistics, Copenhagen, pp 148–156

Dauphin YN, Fan A, Auli M, Grangier D (2017) Language modeling with gated convolutional networks. In: International conference on machine learning. PMLR, pp 933–941

Durrani N, Schmid H, Fraser A (2011) A joint sequence translation model with integrated reordering. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp 1045–1054, Portland

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, pp 489–500

Eriguchi A, Hashimoto K, Tsuruoka Y (2019) Incorporating source-side phrase structures into neural machine translation. Comput Linguist 45(2):267–292

Fadaee M, Bisazza A, 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). Association for Computational Linguistics, Vancouver, pp 567–573

Fadaee M, Monz C (2018) Back-translation sampling by targeting difficult words in neural machine translation. In: EMNLP

Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: Proceedings of the 34th international conference on machine learning-volume 70. JMLR.org, pp 1243–1252

Hassan H, Aue A, Chen C, Chowdhary V, Clark J, Federmann C, Huang X, Junczys-Dowmunt M, Lewis W, Li M, Liu S, Liu T-Y, Luo R, Menezes A, Qin T, Seide F, Tan X, Tian F, Wu L, Wu S, Xia Y, Zhang D, Zhang Z, Zhou M (2018) Achieving human parity on automatic Chinese to English news translation. arXiv:1803.05567

Hieber F, Domhan T, Denkowski M, Vilar D, Sokolov A, Clifton A, Post M (2018) The sockeye neural machine translation toolkit at AMTA 2018. In : Proceedings of the 13th conference of the association for machine translation in the Americas (Volume 1: Research Papers). Association for Machine Translation in the Americas, Boston, pp 200–207

Iyer K (2020) Sentence boundary detection in legal texts grading: Option 3

Jha GN (2010) The TDIL program and the Indian langauge corpora intitiative (ILCI). In: LREC

Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: International conference on learning representation (ICLR)

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

Koehn P (2004) Statistical significance tests for machine translation evaluation. In: Lin D, Wu D (eds), Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP), Barcelona, pp 388–395

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: MT summit. Citeseer, vol 5, pp 79–86

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, et al. (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, Knowles R (2017) Six challenges for neural machine translation. In: Proceedings of the first workshop on neural machine translation. Association for Computational Linguistics, pp 28–39

Koehn P, Och FJ, 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, Bhattacharyya P (2018) The IIT Bombay English-Hindi Parallel Corpus. In: Chair NCC, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T, (eds). Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), European Language Resources Association (ELRA), Paris

Nakazawa T, Higashiyama S, Ding C, Dabre R, Kunchukuttan A, Pa WP, Goto I, Mino H, Sudoh K, Kurohashi S (2018) Overview of the 5th workshop on asian translation. In: Proceedings of the 5th Workshop on Asian Translation (WAT2018), Hong Kong

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. In: ACL-2002: 40th annual meeting of the association for computational linguistics. ACL, Philadelphia, pp 311–318

Ramanathan A, Bhattacharyya P, Visweswariah K, Ladha K, Gandhe A (2011) Clause-based reordering constraints to improve statistical machine translation. In: Proceedings of 5th international joint conference on natural language processing, Philadelphia, pp 1351–1355

Sennrich R, Haddow B, Birch A (2016a) Improving neural machine translation models with monolingual data. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 86–96

Sennrich R, Haddow B, Birch A (2016b) Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, Berlin

Sennrich R, Haddow B, Birch A (2016c) 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, pp 1715–1725

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008

Wang S, Liu Y, Wang C, Luan H, Sun M (2019) Improving back-translation with uncertainty-based confidence estimation. 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, pp 791–802

Wang X, Pham H, Dai Z, 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. Association for Computational Linguistics, Brussels, pp 856–861

Wang X, Tu Z, Xiong D, Zhang M (2017) Translating phrases in neural machine translation. In: Proceedings of the 2017 conference on empirical methods in natural language processing. Association for Computational Linguistics, Copenhagen, pp 1421–1431

Zhang J, 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

Zhao Y, Wang Y, Zhang J, Zong C (2018) Phrase table as recommendation memory for neural machine translation. In: IJCAI

Zhou C, Ma X, Hu J, Neubig G (2019) Handling syntactic divergence in low-resource machine translation. arXiv preprint arXiv:1909.00040

Zhu J, Gao F, Wu L, Xia Y, Qin T, Zhou W, Cheng X, Liu T-Y (2019) Soft contextual data augmentation for neural machine translation. In: ACL

Zoph B, Yuret D, May J, 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, Austin, pp 1568–1575

Authors and Affiliations

Kamal Kumar Gupta[1] · Sukanta Sen[1] · Rejwanul Haque[2] · Asif Ekbal[1] · Pushpak Bhattacharyya[1] · Andy Way[2]

✉ Kamal Kumar Gupta
kamal.pcs17@iitp.ac.in

✉ Asif Ekbal
asif@iitp.ac.in

Sukanta Sen
sukanta.pcs15@iitp.ac.in

Rejwanul Haque
rejwanul.haque@adaptcentre.ie

Pushpak Bhattacharyya
pb@iitp.ac.in

Andy Way
andy.way@adaptcentre.ie

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India

[2] School of Computing, ADAPT Centre, Dublin City University, Dublin, Ireland