Improving Sinhala-Tamil Translation through Deep Learning Techniques

A. Arukgoda[1][0000−0001−5953−9332], A. R. Weerasinghe[2][0000−0002−1392−7791], and R. Pushpananda[3][0000−0001−9082−1280]

[1] University of Colombo School of Computing, Colombo, Sri Lanka
anupama.arukgoda@gmail.com
[2] University of Colombo School of Computing, Colombo, Sri Lanka
arw@ucsc.cmb.ac.lk
[3] University of Colombo School of Computing, Colombo, Sri Lanka
rpn@ucsc.cmb.ac.lk

Abstract. Neural Machine Translation (NMT) is currently the most promising approach for machine translation. But still, due to the data hungry nature of NMT, many of the low resourced language pairs struggle to apply NMT and generate intelligible translations. Additionally, when the language pair is morphologically rich and also when the corpora is multi-domain, the lack of a large parallel corpus becomes a significant barrier. This is because morphologically rich languages inherently have a large vocabulary, and inducing a model for such a large vocabulary requires much more example parallel sentences to learn from. In this research, we investigated translating from and into both a morphologically rich and a low resourced language pair, Sinhala and Tamil, exploring the suitability of different techniques proposed in the literature in the context of Sinhala and Tamil. Through the course of our experiments, we gained a statistically significant improvement of approximately 11 BLEU points for Tamil to Sinhala translation and an improvement of 7 BLEU points for Sinhala to Tamil translation over our baseline systems. In this process we also designed a new language-independent technique that performs well when even the amount of monolingual sentences are limited and could support the translation of one direction on the translation of the other direction, given two languages.

Keywords: Neural Machine Translation (NMT) · Low-resource translation · Sinhala · Tamil.

1 Introduction

Neural Machine Translation (NMT) represents a significant step-forward over a basic statistical approach, therefore is considered as the state-of-the-art for Machine Translation. NMT provides a direct translation mechanism from source language to the target language providing a stronger generalization power whereas in Statistical Machine Translation (SMT) the translation consists of two main stages; translation model and language model, which are trained separately and combined later [3]. Moreover, NMT systems are capable of modeling longer dependencies thanks to the recurrent neural network (RNN) encoder-decoder model. But one of the inherent drawbacks of NMT is that it requires large parallel corpora which limits their applicability to translate under-resourced languages. Sinhala and Tamil are the national languages of Sri Lanka. Sinhala and Tamil are both morphologically rich, they are low resourced (limited number of parallel corpora available) and have limited or no publicly available linguistic resources such as POS taggers and morphological analyzers which could have supported the translation. These properties make the task of translating Sinhala and Tamil more challenging. In an early research in SMT for Sinhala and Tamil translation [18], it has been shown that due to the co-evolution of the Sinhalese and Tamils in Sri Lanka, the linguistic distance between Sinhala and Tamil is less than that between Sinhala and English, thereby making the translation between Sinhala and Tamil theoretically easier than that of Sinhala and English. In addition, the two languages Sinhala and Tamil, are syntactically-similar, which also provide the flexibility to alter the word order. These commonalities between Sinhala and Tamil increase the feasibility of their translation. The currently available best open-domain Sinhala-Tamil translator is based on SMT [8](will be referred to as the SMT study). Their work has been conducted for only Tamil to Sinhala translation direction. We have employed the exact same parallel corpus used in the SMT study to make a fair comparison between the performance of SMT versus NMT on the same parallel corpus.

This paper has two main contributions: By identifying the two main challenges of translating the two languages under consideration, we explored suitable techniques to treat them and the observed results were compared with the accuracy reported by the SMT study on the exact same parallel corpus. Such a detailed analysis to improve the open-domain translation of Sinhala and Tamil using NMT have not been conducted before, to the best of our knowledge. Secondly, an observation throughout the research was that Tamil to Sinhala translation performs better than Sinhala to Tamil translation which prompted us to propose a more general-purpose, language-independent method which can make use of the accuracy improvement in one translation direction on the improvement of the other translation direction which can be explored and adopted by other such low-resourced languages.

2 Literature Review

2.1 Neural Machine Translation

Machine translation is a sequence prediction problem. Not only both input and output are sequences, they are sequences of different lengths which made the task more challenging. But the pioneering work of [14] presented an end-to-end sequence learning approach that makes minimal assumptions on the sequence length and structure, outperforming the traditional phrase-based translation systems. The most popular architecture for NMT is the encoder-decoder architecture. This model predicts a target word, based on the context vectors associated with the source positions and all the previous generated target words. This is the traditional architecture used in NMT and since its introduction much research work have been conducted to improve this architecture [1].

2.2 Translating Morphologically Rich Languages

A Morphologically Rich Language (MRL) is one which grammatical relations such as tense, singularity/plurality, predicate, gender, age etc. [15], are indicated by changes to the words instead of relative position or addition of particles. Dealing with morphologically rich languages is an open problem in language processing as the complexity in the word forms inherent to these languages makes translation complex. Most of the machine translation systems are trained using a fixed vocabulary. But translation is an open vocabulary problem. Therefore, having to deal with out-of-vocabulary (OOV) words, and rare words is unavoidable. If the translated languages are low-resourced (size of the parallel corpora is small) and the corpora are multi-domain, this problem is worsened because of the increased size of the vocabulary and the increased amount of word senses. Hence, translation mechanisms that go below the word level such as Byte-Pair-Encoding (BPE)[12] and Morfessor[13] have been proposed.

2.3 Translating Low-Resourced Languages

Similar to many other deep learning tasks, the success of NMT is strongly dependent on the availability of large parallel corpora. Since this is a luxury many of the languages (especially minority languages) do not have, many techniques have been proposed over the years to address this. One such technique is to incorporate monolingual corpora by translating monolingual corpora with a translator trained in the backward direction and thereby create synthetic parallel sentences making the overall parallel corpus size larger [21, 10]. The intuition is that even though it is quite difficult to obtain a large parallel corpus for two languages, it is much easier to obtain large monolingual corpora for the two languages separately. This technique has been applied for back-translation of both source-side monolingual corpora [21] and target-side monolingual corpora [10]. While this paved way to improve the translation quality of low-resourced languages making maximum use of both parallel and monolingual corpora, it has also been shown empirically that such models tend to 'forget' the correct semantics of translation if trained on much more synthetic data than authentic parallel data [7], imposing a constraint on the amount of monolingual data that can be used.

Another reason for the popularity of the back-translation technique was it required no changes to the network architecture. Therefore, techniques have been introduced to improve the quality of the back-translator, since it is yet another imperfect MT system. Imankulova et al. [5], proposes a filtering technique which chooses the back-translated synthetic sentences with the highest quality. This had improved the final translation quality leading to higher BLEU scores.

2.4 Sinhala - Tamil Translation

The best Sinhala-Tamil translator to-date has been produced through the most recent research for morphologically rich languages [8], based on statistical machine translation. The authors have integrated an unsupervised morphological modification approach called Morfessor, suggested in a previous research on Sinhala morphological analysis [19] to overcome the issues related to morphological richness. This has resulted in dramatic improvements in the translation quality and the reliability of the Sinhala - Tamil translation.

3 Methodology

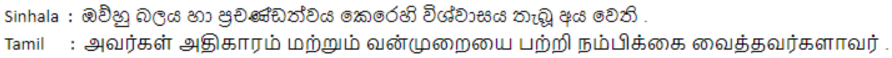
In this research, we first treated the first challenge for translating between Sinhala and Tamil, the morphological richness of both languages by exploring the impact of two sub-word segmentation techniques and thereby reduced the size of the vocabulary of the corpus and treated the OOV problem. Next, we explored the suitability of two back-translation techniques using our open-domain monolingual corpora to increase the corpus size and thereby treated the second main challenge for the translation of Sinhala and Tamil and improved the translation accuracy further. We constrained ourselves to mainly explore back-translation techniques as they can make the maximum use of our open-domain parallel and monolingual corpora. Since our baseline SMT study has made use of the same corpora, we were able to make a fair comparison between SMT and NMT in the context of Sinhala and Tamil this way. Finally we proposed a novel technique called Incrementally Filtered Back-Translation and explored its impact on the translation of Sinhala and Tamil.

4 EXPERIMENTAL SETUP

4.1 Data-set Details

For our experiments we used a parallel corpus of approximately 25000 sentences which have a sentence length between 8 and 12 words, collected in the research [8]. Figure 1 shows an example parallel sentence from this corpus. When exploring back-translation techniques to improve Sinhala-Tamil translation accuracy, a 10-million-word monolingual corpus [16] and on the Tamil end, a 4.3-million-word Sri Lankan Tamil monolingual corpus [17] were used. Out of these original monolingual corpora, sentences having a length of 8 to 12 words were extracted for our work. Both these corpora are suitable for an open-domain translation as they have been collected from sentences from different domains such as newspaper articles, technical writing and creative writing. The corpus statistics for the parallel and monolingual corpora are provided in Tables 1 and 2 respectively.

Fig. 1. An example Sinhala and Tamil parallel sentence pair



Sinhala : ඔව්හු බලය හා ප්‍රචණ්ඩත්වය කෙරෙහි විශ්වාසය තැබූ අය වෙති .

Tamil : அவர்கள் அதிகாரம் மற்றும் வன்முறையை பற்றி நம்பிக்கை வைத்தவர்களாவர் .

Table 1. Parallel Corpus Statistics

Corpus Statistics	**Sinhala**	**Tamil**
Sentence Pairs	26,187	
Vocabulary Size (V)	38,203	54,543
Total number of words (T)	262,082	227,486
V/T %	14.58	23.98

4.2 Pre-Processing

We first obtained three different representations of our original corpora. The first one being the original full word-form corpus (i.e the corpus as it is without any pre-processing) and the second being the corpora segmented into morpheme-like units. For this segmentation, we used the tool Morfessor 2.0 which provides a morpheme segmentation algorithm that works in an unsupervised manner and aims to generate the most probable segmentation of words to their prefix, suffix and stem by relying on the Minimum Description Length from the words in an un-annotated/raw corpus. Figure 2 shows a few example outputs from Morfessor 2.0.

In order to make the post-processing easier after translation of the sub-words, we introduced a special character '@@' between the sub-words such that the boundary of the words are maintained. Then, after the translation, the sub-words can be put back to the original word form by merging each sub-word with the special character with the immediately next sub-word.

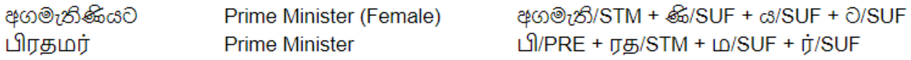
The third form of representation was obtained by pre-processing the full-word corpora using the algorithm Byte-Pair-Encoding (BPE). This algorithm requires the tuning of the number of merge operations, which is an input parameter that solely depends on the language and the data-set. We empirically chose a value of 750 for Sinhala to Tamil translation direction and a value of 1000 for Tamil to Sinhala translation as the number of merge operations.

Figure 3 shows an example Sinhala sentence pre-processed with each technique.

Table 2. Monolingual Corpus Statistics

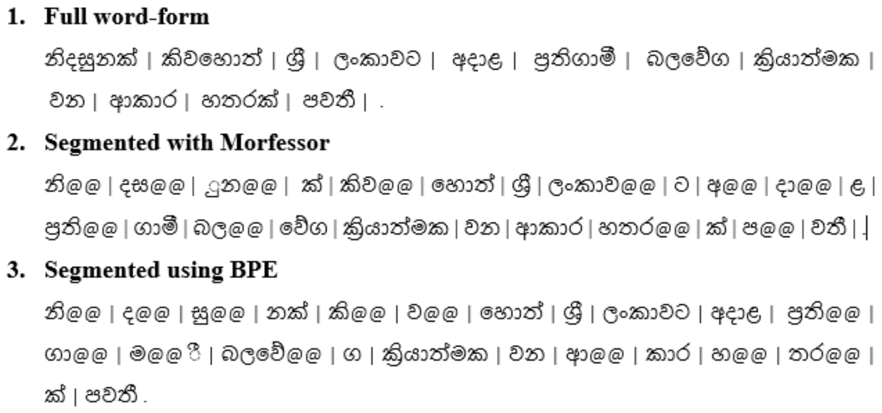
Corpus Statistics	**Sinhala**	**Tamil**
Number of Sentences	180,793	40,453
Vocabulary Size	154,782	65,228
Total number of words	1,577,921	352,813

Fig. 2. Examples of unsupervised morphological decomposition with Morfessor 2.0



අගමැතිණියට	Prime Minister (Female)	අගමැති/STM + ණි/SUF + ය/SUF + ට/SUF
பிரதமர்	Prime Minister	பி/PRE + ரத/STM + ம/SUF + ர்/SUF

Fig. 3. The three pre-processing techniques



1. **Full word-form**

   නිදසුනක් | කිවහොත් | ශ්‍රී | ලංකාවට | අදාළ | ප්‍රතිගාමී | බලවේග | ක්‍රියාත්මක | වන | ආකාර | හතරක් | පවතී | .

2. **Segmented with Morfessor**

   නි@@ | දස@@ | ුන@@ | ක් | කිව@@ | හොත් | ශ්‍රී | ලංකාව@@ | ට | අ@@ | දා@@ | ළ | ප්‍රති@@ | ගාමී | බල@@ | වේග | ක්‍රියාත්මක | වන | ආකාර | හතර@@ | ක් | ප@@ | වතී | .

3. **Segmented using BPE**

   නි@@ | ද@@ | සු@@ | නක් | කි@@ | ව@@ | හොත් | ශ්‍රී | ලංකාවට | අදාළ | ප්‍රති@@ | ගා@@ | ම@@ ී | බලවේ@@ | ග | ක්‍රියාත්මක | වන | ආ@@ | කාර | හ@@ | තර@@ | ක් | පවතී .

4.3 Baseline Model

In order to compare our results we initiated two benchmarks. The first benchmark is the translation accuracy reported for Tamil to Sinhala translation with SMT in the work of [8]. We used the exact same parallel corpus as given in the SMT study, to see the performance of NMT in comparison to SMT with the same parallel corpus.

The SMT study has been conducted only in the direction of Tamil to Sinhala. They have reported a translation accuracy of 13.11 BLEU points. Since in our research work we are interested in the translation in both Sinhala to Tamil and Tamil to Sinhala directions, we used a second baseline model by training a network with the architecture provided under System Setup on our 25000 full-word form parallel corpus.

4.4 System Setup

We used the framework OpenNMT [6] for the experiments. We initiated our experiments with a 2-layer Bidirectional Recurrent Neural Network (BRNN) with 500 hidden units on both encoder and decoder. To speed-up the training process GeForce GC 1080 Ti GPU was used with a GPU memory of 16 GB. The sole indication of the translation accuracy throughout the experiments was Bilingual Evaluation Understudy (BLEU) score. The reported BLEU scores were obtained by performing 3-fold cross validation to reduce any biases.

4.5 Manipulating the Network

Simplifying the network After translating the three differently pre-processed parallel corpora, we chose the best pre-processing technique out of them. Next we simplified the network by using a Google Neural Machine Translate (GNMT) encoder [20] instead of a BRNN encoder which was being used in the experiments so far. A BRNN encoder has bidirectional connections (to process each sentence from left to right as well as from right to left) between the neurons in each layer, whereas in the GNMT encoder only the first layer is a single bidirectional layer and the other layers are unidirectional RNN layers. The bidirectional states in this layer are concatenated and residual connections are fed to the next layers which are uni-directional.

Checkpoint Smoothing We went a step further to improve the BLEU scores, and that is by using an ensemble technique. Traditionally, ensemble methods are learning algorithms that combine multiple individual methods to create a learning algorithm that is better than any of the individual parts. Checkpoint smoothing is one such ensemble technique which uses only a single training process [4]. The idea is, rather than using the model generated from the final epoch, we average the parameters of the models from multiple epochs and translate using the averaged models. Such an averaged model is expected to produce better translations.

4.6 Back-Translation

Naive Back-Translation As proposed in [10] we trained a back-translator using the authentic parallel sentences and iteratively added synthetic parallel sentences obtained by translating randomly selected target-side monolingual sentences until the ratio between authentic and synthetic parallel sentences is 1:3 for Tamil to Sinhala translation and the same value is 1:2 for Sinhala to Tamil translation.

Filtered Back-Translation As proposed by Imankulova et al. [5] we filtered the synthetic parallel sentences obtained by translating the target-side monolingual corpus, using sentence-level similarity metric BLEU score and iteratively added the synthetic parallel sentences with the best BLEU score until the ratio between authentic and synthetic parallel sentences is 1:3 for Tamil to Sinhala translation and the same value is 1:2 for Sinhala to Tamil translation.

Throughout the experiments we observed that Tamil to Sinhala translation direction performed significantly better than Sinhala to Tamil translation direction. This encouraged us to design a technique which can use the improvement in one translation direction on the accuracy improvement in the other translation direction and vice verse, given two languages. As a solution we introduced the algorithm presented as Incrementally Filtered Back-Translation shown under Algorithm 1.

Algorithm 1: Incrementally Filtered Back-Translation

```
Input: authentic parallel sentences (auth-parallel), monolingual sentences
       from language-1 ($mono_{lang1}$), monolingual sentences from language-2
       ($mono_{lang2}$), k=1
Let src = language-1
Let tgt = language-2
Let $\theta_{\rightarrow}$ = model trained from src to tgt with auth-parallel
Let $\theta_{\leftarrow}$ = model trained from tgt to src with auth-parallel
Let D = auth-parallel
repeat
    filtered-synthetic-parallel = Filter($\theta_{\rightarrow}$ , $\theta_{\leftarrow}$) ;  // Call Filter algorithm
      provided in Algorithm 2
    D = D $\cup$ filtered-synthetic-parallel ;                // Combine the
      filtered-synthetic parallel sentences with the authentic
      parallel corpus
    $\theta_{\rightarrow} = \theta_{\leftarrow}$
    $\theta_{new}$ = Model trained on D from src to tgt
    $\theta_{\leftarrow} = \theta_{new}$
    src = language-2
    tgt = language1
until convergence-condition || $\big( |mono_{tgt}| = 0 \big)$;
return Newly updated model $\theta_{new}$
```

4.7 Incrementally Filtered Back-Translation

With this new algorithm the translation in both translation directions are done in parallel. The first step is similar to filtered back-translation technique. But unlike filtered back-translation which used two models trained in each direction using the authentic parallel corpus, with this technique, the updated models in each iteration are used to create and filter the synthetic parallel sentences such that the quality of the synthetic parallel sentence added in each iteration are of much better quality. The translation accuracy improvement obtained in one translation direction is used to improve the translation in the other translation direction.

The algorithm executes until the improvement in the BLEU score on the same test-set on models from two consecutive iterations is insignificant (convergence-condition) or if all the monolingual sentences are consumed in target-language monolingual corpus.

5 Evaluation

The BLEU scores reported by the three representation forms are shown in Table 3. The initial results from the full-word form representation for Tamil to Sinhala translation direction was 5.41 and Sinhala to Tamil was 2.47, which were discouraging. The translations did not resemble the semantics of the reference

```
Algorithm 2: Filter($\theta_{\rightarrow}$ , $\theta_{\leftarrow}$)
Input: /* Assume all the variables are being shared between Algorithm 1 and
        Algorithm 2 */
Get synthetic src sentences (synth-src) by translating $mono_{tgt}$ with $\theta_{\leftarrow}$
Get synthetic tgt sentences (synth-tgt) by translating synth-src with $\theta_{\rightarrow}$
BLEU($mono_{tgt}$, synth-tgt) ;        // Calculate BLEU score by comparing
  synth-tgt against $mono_{tgt}$
Sort $mono_{tgt}$ in descending order of the BLEU score
Choose the first x amount of $mono_{tgt}$ sentences (x-$mono_{tgt}$) and the
  corresponding synthetic source language sentences (x-synth-src) such that
  the ratio between authentic parallel sentences to synthetic sentences is 1:k
Create a pseudo-parallel corpus S = { x-synth-src, x-$mono_{tgt}$ }
$mono_{tgt}$ = $mono_{tgt}$ - (x-$mono_{tgt}$) ;    // Update $mono_{tgt}$ by removing the
  chosen top-x mono sentences from $mono_{tgt}$
k = $k + 1$
return S
```

sentences. Only a few words from each sentence were correctly translated but these did not add any value to the underlying meaning of the sentence.

As can be seen in Table 3, translations with both sub-word segmentation have performed drastically better than the baseline full-word form. With Morfessor, the improvement is 3.76 for Tamil to Sinhala and 3.59 for Sinhala to Tamil. Compared to the full-word form BPE representation has improved the BLEU score by 4.6 BLEU points for Tamil to Sinhala and by 3.94 for Sinhala to Tamil. To justify this improvement we analyzed the OOV% of the parallel corpus obtained with each representation. This analysis is provided in Table 4. The OOV% has dropped drastically with sub-word segmentation techniques. This means that sub-word segmentation increase the coverage of the model, enhancing the amount of words a model can see during the training. This has positively affected the translation resulting in increased BLEU scores.

Another observation from Table 3 is that BPE performs better than Morfessor for Sinhala and Tamil. One reason could be as we see in Table 4, BPE has decreased the OOV% much more than with Morfessor. Also we observed that when pre-processed with Morfessor, more words were segmented into better linguistic morphemes than with BPE (as can also be witnessed in Figure

Table 3. BLEU scores of the three pre-processing techniques

Representation	Tamil-Sinhala	Sinhala-Tamil
SMT (Baseline 1)	13.11	-
Full-word form (Baseline 2)	5.41	2.47
Morfessor	9.17	6.06
BPE	10.01	6.41

Table 4. OOV% Analysis

Representation	**Tamil-Sinhala OOV%**	**Sinhala-Tamil OOV%**
Full-word form	34.46	24.54
Morfessor	6.27	2.60
BPE	0	0

Table 5. BLEU Scores after manipulating the network

Technique	**Tamil-Sinhala**	**Sinhala-Tamil**
GNMT Encoder	10.57	6.94
+Checkpoint Smoothing	11.76	7.51

3). Similar observations were made in the work of Banerjee et al. [2] on the Bengali-Hindi pair of languages. They empirically show in their work that for linguistically close languages, BPE performs better than when using Morfessor. The above conclusions help us derive another conclusion. That is, since linguistically similar languages like Sinhala and Tamil benefit from the fact that the sub-units are not segmented into proper morphemes, it frees us from the need of a morphological analyzer for NMT translation tasks. Therefore we could focus our future efforts on improving the quality of the newly generated words when translated after pre-processing with BPE by incorporating a Language Model.

The next experiments were continued by using the pre-processing technique BPE.

When we simplified the network by replacing the BRNN encoder with a GNMT encoder, the translation accuracy improved further as shown in Table 5, by 0.56 for Tamil to Sinhala and 0.53 for Sinhala to Tamil. We noticed that the number of parameters used in the model with a BRNN encoder (22,683,128) was almost as twice as the number of parameters of that with a GNMT encoder (11,104,741). When the data-set is small, we cannot afford to fit models with a high degree of freedom (too many parameters). This leads to the requirement of a simpler model which has led to the improvement in the translation quality.

Furthermore the ensemble of models from multiple epochs used to create the averaged model has increased the generalization power of the models, resulting in a significant improvement of 1.19 BLEU points for Tamil to Sinhala and 0.57 for Sinhala to Tamil in the translation accuracy.

The series of experiments so far were conducted only using the parallel corpus of approximately 25000 sentences. In the following experiments we attempted to make use of our monolingual corpora to increase the net parallel corpus size by using back-translation techniques. The results from naive and filtered back-translation techniques are presented in the tables 6 and 7 respectively. As the back-translators we used the best models we obtained with the authentic 25000 parallel sentences after employing Checkpoint Smoothing (Table 5).

Table 6. BLEU Scores from Naive Back-Translation

Authentic : Synthetic	**Tamil-Sinhala**	**Sinhala-Tamil**
1 : 1	12.16	7.34
1 : 2	14.17	7.37
1 : 3	15.35	-

Table 7. BLEU Scores from Filtered Back-Translation

Authentic : Synthetic	**Tamil-Sinhala**	**Sinhala-Tamil**
1 : 1	14.04	7.23
1 : 2	14.75	7.58
1 : 3	15.93	-

The Tamil to Sinhala translation direction has continuously improved when the ratio between authentic to synthetic parallel sentences were increased. And as expected, filtered back-translation performs better than naive back-translation as filtered translation ensures that even the synthetic parallel sentences are of good quality, unlike naive back-translation. The observations conformed to the common wisdom of "more data is better data" in the context of Deep Learning. But the expected improvement in the translation quality through these techniques were not witnessed in the translations conducted from Sinhala to Tamil which questions their applicability across languages.

One of the consistent observations in our research work and previous work [11, 9] is that given two languages, translation in one direction performs better than the other. This distinction is more prominent when one language is morphologically richer than the other. This prompted us to design an algorithm that benefits from this fact and improve the quality of both translation directions. Our algorithm, also known as 'Incrementally Filtered Back-Translation', manages to help the translations reach high accuracy with minimum amount of monolingual sentences. This is an original contribution by us to the body of knowledge. The results from this newly proposed technique is given in Table 8. As expected, this new technique was able to create better translation accuracy for both translation directions (shown in Table 8). Sinhala to Tamil direction had increased its

Table 8. BLEU Scores from Incrementally Filtered Back-Translation

Authentic : Synthetic	**Tamil-Sinhala**	**Sinhala-Tamil**
1 : 1	14.04	-
1 : 2	-	9.41
1 : 3	15.39	-
1 : 4	-	9.71
1 : 5	16.02	-

BLEU score by approximately 2 points more than the accuracy obtained with only the parallel corpus, which is a significant improvement. The importance in the technique is that, the improvement in the BLEU scores are seen at the earlier stages. When the ratio between authentic to synthetic parallel sentences was 1:1, Sinhala to Tamil translation direction obtained a BLEU score of 9.41 with the newly proposed Incrementally Filtered Back-Translation, while the same value was only 7.23 with filtered back-translation. This makes this technique ideal when even the amount of monolingual sentences available for two languages is limited as it makes maximum use of even the limited amount of monolingual sentences. Furthermore, since this technique is language-independent, it can be adopted and explored on any such low-resourced language pair.

The final BLEU scores achieved by our study are compared against the benchmarks in Table 9. We were also able to exceed the benchmark based on the SMT study for Tamil to Sinhala translation direction by approximately 3 BLEU points. While the same parallel corpus was used in the two studies, the SMT study had used 850,000 Sinhala monolingual sentences for its language model while the Sinhala monolingual corpus we used for our work consisted of only 180,793 sentences. Therefore by using less amount of resources, NMT had managed to exceed the translation accuracy more than SMT. An observation we made throughout the experiments was that the translation of Tamil to Sinhala performed better than the translation for Sinhala to Tamil. If we consider the characteristics of the Sinhala and Tamil parallel data-sets in Table 1, we can clearly see that Vocabulary to Total Words ratio of the Tamil data-set is 23.98% which is almost two times larger than the same value for Sinhala data-set, which is 14.58%. This is an indication that Tamil is morphologically richer than Sinhala within our corpora. Also it can be observed from Table 1, Sinhala has a higher number of total words than Tamil. Since in the parallel corpus, the sentences of the two languages have the exact meaning of each other, it can be stated that Sinhala requires more words than Tamil to be used to convey the same meaning. When a language is morphologically richer, the inflectional morphemes add more information about time, count, singularity/plurality etc. Therefore a morphologically richer language requires only fewer words to convey a message than a relatively less morphologically rich language. Therefore we conclude that within the context of our corpora, Tamil behaves morphologically richer than Sinhala.

NMT is an end-to-end translation. In an encoder-decoder architecture, the encoder encodes a source sentence in an almost language independent representation which will later be decoded on the decoder-side. When the source-side is

Table 9. Comparison of final BLEU Scores against the Baseline Models

Model	**Tamil-Sinhala**	**Sinhala-Tamil**
SMT (Baseline 1)	13.11	-
Full-word form (Baseline 2)	5.41	2.47
Incrementally Filtered BT	16.02	9.71

morphologically richer than the target-side, the encoder tends to encode more information about the sentence, leading to a better decoding by the decoder. When the source-language is less morphologically rich than the target-side, the encoded sentences does not contain much information for the decoder to deduce a good translation. This justifies why Tamil to Sinhala translation direction produces better translations than for Sinhala to Tamil translations.

6 Conclusion

Our main goal in this research was to develop an NMT system to improve the translation between the morphologically rich and low resource language pair Sinhala and Tamil. By identifying the challenging properties of Sinhala and Tamil and by treating them appropriately through a course of experiments, we improved the NMT benchmark by a BLEU score of 11 for Tamil to Sinhala direction, and 7 for Sinhala to Tamil translation direction. Using the same parallel corpus as the SMT study we were able to exceed the SMT benchmark for Tamil to Sinhala translation direction by approximately 3 BLEU points.

This research paves way for the newly proposed Incrementally Filtered Back-Translation technique to be explored by other low resource languages and establish its validity across languages. Furthermore, the techniques that have been accepted as more suitable for Sinhala and Tamil translation can be explored and adopted by other such agglutinative languages as well. While we hope that this research contributes to the improvement of the information exchange between Sinhala and Tamil communities, we have successfully addressed the gap in the body of knowledge as research on open-domain Sinhala Tamil using NMT has not been attempted before.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014), http://arxiv.org/abs/1409.0473
2. Banerjee, T., Bhattacharyya, P.: Meaningless yet meaningful: Morphology grounded subword-level nmt. In: Proceedings of the Second Workshop on Subword/Character LEvel Models. pp. 55–60. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-1207, http://aclweb.org/anthology/W18-1207
3. Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M.: Neural versus phrase-based machine translation quality: a case study. CoRR **abs/1608.04631** (2016), http://arxiv.org/abs/1608.04631
4. Chen, H., Lundberg, S., Lee, S.: Checkpoint ensembles: Ensemble methods from a single training process. CoRR **abs/1710.03282** (2017)
5. Imankulova, A., Sato, T., Komachi, M.: Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In: WAT@IJCNLP. pp. 70–78. Asian Federation of Natural Language Processing (2017)

6. Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., Rush, A.M.: Opennmt: Neural machine translation toolkit. In: AMTA (1). pp. 177–184. Association for Machine Translation in the Americas (2018)
7. Poncelas, A., Shterionov, D., Way, A., de Buy Wenniger, G.M., Passban, P.: Investigating backtranslation in neural machine translation (2018)
8. Pushpananda, R., Weerasinghe, R., Niranjan, M.: Statistical machine translation from and into morphologically rich and low resourced languages (04 2015)
9. Sennrich, R., Birch, A., Currey, A., Germann, U., Haddow, B., Heafield, K., Barone, A.V.M., Williams, P.: The university of edinburgh's neural MT systems for WMT17. CoRR **abs/1708.00726** (2017)
10. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data (11 2015)
11. Sennrich, R., Haddow, B., Birch, A.: Edinburgh neural machine translation systems for WMT 16. CoRR **abs/1606.02891** (2016)
12. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics (2016), http://aclweb.org/anthology/P/P16/P16-1162.pdf
13. Smit, P., Virpioja, S., Grönroos, S., Kurimo, M.: Morfessor 2.0: Toolkit for statistical morphological segmentation. In: Bouma, G., Parmentier, Y. (eds.) Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden. pp. 21–24. The Association for Computer Linguistics (2014), http://aclweb.org/anthology/E/E14/E14-2006.pdf
14. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014), https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
15. Vylomova, E., Cohn, T., He, X., Haffari, G.: Word representation models for morphologically rich languages in neural machine translation. In: Proceedings of the First Workshop on Subword and Character Level Models in NLP. pp. 103–108. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/W17-4115, https://www.aclweb.org/anthology/W17-4115
16. Weerasinghe, R., Herath, D., Welgama, V., Medagoda, N., Wasala, A., Jayalatharachchi, E.: Ucsc sinhala corpus - pan localization project-phase i (2007)
17. Weerasinghe, R., Pushpananda, R., Udalamatta, N.: Sri lankan tamil corpus. Technical report, University of Colombo School of Computing and funded by ICT Agency, Sri Lanka, (2013)
18. Weerasinghe, R.: A statistical machine translation approach to sinhala tamil language translation. In: In SCALLA 2004 (2004)
19. Welgama, V., Weerasinghe, R., Niranjan, M.: Evaluating a machine learning approach to sinhala morphological analysis (12 2013)
20. 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, L., 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., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016)

21. Zhang, J., Zong, C.: 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. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/D16-1160, http://aclweb.org/anthology/D16-1160