Handling Rare Word Problem using Synthetic Training Data for Sinhala and Tamil Neural Machine Translation

Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, Surangika Ranathunga
Department of Computer Science and Engineering
University of Moratuwa
Katubedda 10400, Sri Lanka
{pasindu.13, prabath.sandaruwan.13, malith.13, narmada.ah.13,surangika,}@cse.mrt.ac.lk

Abstract

Lack of parallel training data influences the rare word problem in Neural Machine Translation (NMT) systems, particularly for under-resourced languages. Using synthetic parallel training data (data augmentation) is a promising approach to handle the rare word problem. Previously proposed methods for data augmentation do not consider language syntax when generating synthetic training data. This leads to generation of sentences that lower the overall quality of parallel training data. In this paper, we discuss the suitability of using Parts of Speech (POS) tagging and morphological analysis as syntactic features to prune the generated synthetic sentence pairs that do not adhere to language syntax. Our models show an overall 2.16 and 5.00 BLEU score gains over our benchmark Sinhala to Tamil and Tamil to Sinhala translation systems, respectively. Although we focus on Sinhala and Tamil NMT for the domain of official government documents, we believe that these synthetic data pruning techniques can be generalized to any language pair.

Keywords: Neural Machine Translation, POS Tagging, Morphological Analysis, Data Augmentation

1. Introduction

Neural Machine Translation (NMT) is the current state-of-the-art machine translation architecture that aims at building a single neural network that can be jointly tuned to maximize the translation performance (Bahdanau et al, 2014). Despite being successful in producing acceptable outputs for language pairs having large parallel corpora (Sennrich et al, 2016), NMT performs poorly for language pairs that lack the luxury of having sufficiently large parallel data (Tennage et al, 2017). This is due to the requirement of numerous instances of sentence pairs with words occurring in different contexts in order to accurately train a NMT model. Being an under-resourced language pair that is unable to satisfy this requirement, Sinhala and Tamil NMT falls short of reaching state-of-the-art performances (Tennage et al, 2017).

Limited corpus size directly influences the rare word problem. Rare word problem refers to the inability of the neural network to properly model words that appear in the corpus only a very few times. Being morphologically rich languages, there exist many inflections for each word in Sinhala and Tamil languages. Hence having many rare words in the corpus is inevitable.

One way to handle the rare word problem is to increase corpus size using synthetic parallel training data. To generate synthetic data, Fadaee et al. (2017) presented a possible technique. This creates new contexts for rare words when generating synthetic training data, thus giving a possible solution for the rare word problem. However, the main limitation of this approach is that this synthetic sentence pair generation technique does not take language syntax into consideration, which eventually lowers expected BLEU score gain.

In this paper, we present two sentence pruning techniques based on Parts of Speech (POS) tagging and morphological analysis to remove synthetic sentence pairs that do not preserve language syntax. Compared to Fadaee et al.'s (2017) method, POS tagging method shows an improvement of 1.04 and 2.12 BLEU score gains for Sinhala to Tamil (Si-Ta) and Tamil to Sinhala (Ta-Si) models, respectively. Use of morphological analysis improves the quality of translation by 1.26 and 2.98 BLEU scores, respectively. Overall, synthetic parallel training data methods yield an improvement of 2.16 and 5.00 BLEU score gains over our benchmark Si-Ta and Ta-Si models.

2. Background and Related Work

2.1 Sinhala and Tamil Languages

Sinhala language descends from Indic language family and Tamil from Dravidian family (Pushpananda et al, 2014). Being morphologically rich, Sinhala has up to 110 noun word forms and up to 282 verb word forms and Tamil has 40 noun word forms and up to 240 verb word forms. Both these languages have the same word order of Subject-Object-Verb.

Tennage et al, (2017) have built the first NMT system for this language pair. Lack of language resources and data sparseness that is caused by morphological variances have been identified as the key factors that hinder the translation performance (Tennage et al, 2017).

2.2 Neural Machine Translation

NMT is an end-to-end translation process that treats a word as the smallest unit (Bahdanau et al, 2014). Encoder Decoder (ED) architecture with attention mechanism is the current state-of-the-art NMT architecture (Cho et al, 2014). In the ED architecture, recurrent activation function is applied recursively over the input sequence until the end when the final internal state of the recurrent neural network (RNN) contains the summary of the whole input sentence. Decoder computes RNN's internal state based on the summary vector, the previous predicted word, and the previous internal state. Using internal hidden state of the decoder, it is possible to score each target word based on how likely it is to follow all the preceding translated words. Using softmax normalization, scores are turned into probabilities.

2.3 Data Augmentation

Fadaee et al. (2017) presented a data augmentation approach that targets low-frequency words by generating new sentence pairs containing rare words in new, synthetically created contexts. They have produced experimental results on low-resource settings and have achieved considerable improvement over the benchmark systems. They have focused mainly on fluency and grammatical structure of synthetic training data, and have disregarded its syntax correctness.

Strategies to train with monolingual data without changing the neural network architecture have been proposed by Sennrich et al. (2016). It is based on the intuition that encoder-decoder NMT architecture already has the capacity to learn the same information as a language model. By pairing monolingual training data with an automatic back-translation, synthetic parallel training data are generated. Quality of synthetic training data generated using this method highly depends on the machine translator that is used for back translation.

3. Methodology

3.1 Initial Data Augmentation

For initial synthetic sentence generation, we use the technique used by Fadaee et al. (2017). Initially we obtain a list of rare words by considering the unique words and their counts. Words that appear only *R* (rare word threshold) times or less are considered as rare words. For each rare word *r*, we iterate through each sentence pair in our parallel corpus.

In the below expressions, s_i and t_i denote the i^{th} word in the source sentence and target sentence, respectively. Each word in the source sentence is iterated through and substituted by *r*. Trigram language probability around *r* is checked thereafter.

If the i^{th} source word is substituted,

Original language probability p_1 = LM (s_{i-1}, s_i, s_{i+1})
Synthetic language probability p_2= LM (s_{i-1}, r, s_{i+1})

if ($p_2 > M * p_1$): this is a valid source substitution (M is fluency threshold).

To generate the target side synthetic sentence, we need to substitute the translation of r to the word in the corresponding original target sentence that is aligned to the word that we removed from the source sentence. Statistical approach of automatic word alignment (Och, Ney, 2004) is used to accomplish this task. Using automatic word alignment, it is possible to get the index of the target word that is aligned with the source word that was removed.

To get the translation of a rare word *r*, phrase tables that are generated using word alignment are used. For a given word *e*, there exist several possible translations *f* according to the generated phrase tables. To find the exact translation, we use a two-way translation probability as follows.

translation(e) = argmax $_{f \in \text{possible translations}}$ (p(f|e)*p(e|f))

where,

p(f|e) : Probability of *f* being the translation of *e*.
p(e|f) : Probability of *e* being the translation of *f*.

If there exists a target side word *q* corresponding to *r*, with two-way translation probability greater than *T* (translation threshold), we select it as a viable translation for *r*. *q* is substituted to the word that is aligned to the word that was removed in source side. If the trigram language probability around that word is greater than *M* times the original trigram language probability, then we select it as a correct target word substitution.

A synthetic sentence pair that satisfies all these conditions is added to the synthetic parallel corpus. To reduce distortion of the meaning, only a single rare word substitution per sentence was allowed. Use of language modeling ensures the fluency of synthetic sentences whereas the use of the translation modeling ensures the correspondence between source sentence and target sentence. Table 1 depicts an example synthetic sentence pair.

Original Sentence Pair	**Synthetic Sentence Pair**
එසේ පවරා දෙනු ලැබුවේ කවරෙකුටද (/esea pavaraa denu lAbuwea kavarekuTada/) - (It was assigned to whom?)	එසේ පවරා දෙනු ලැබුවේ ඔබටද? (/esea pavaraa denu labuwea obaTada /) (It was assigned to you?)
அவ்வாறு யாருக்கு ஒப்படைக்கப்பட்டுள்ளது? (/avvaaRu yaarukku oppataikkappattuLLadhu/) (It was assigned to whom?)	அவ்வாறு யாருக்கு உங்களுக்கும்? (/avvaaRu yaarukku ungkaLukkum/) (It was to whom and to you)[1]

Table 1: Initially Generated Synthetic Sentence Pair

Human evaluation of the synthetic parallel training data generated using this method revealed that the resulting sentences do not preserve language syntax. Hence, we investigated on methods to prune the synthetic sentence pairs that do not preserve language semantics.

3.2 Parts of Speech Tagging

POS tags contain important syntactic information about the word in the context that it appears. Based on this property, we further increased the quality of synthetic training data by checking the POS tag of each rare word that is substituted.

Initially, the original parallel corpus is POS tagged. Then using the methodology proposed in section 3.1, all possible synthetic sentence pairs are generated. Then the synthetic parallel sentences are also POS tagged. Algorithm 1 describes this method.

Here,
s_i= word that was removed from source sentence.
t_i= word that was removed from target sentence.
r= rare word that was introduced to source sentence.
t= translation of r that was introduced to target side.

Algorithm 1 – POS tag based pruning

```
if (POS tag of $s_i$ == POS tag of $r$) and (POS tag of $t_i$ == POS tag of $t$) then
    Keep the synthetic sentence pair
else
    Remove it from the corpus
```

[1] Word ordering in Sinhala is different from English. The exact English translations are “To whom was it assigned?” / “Was it assigned to you?” (“To whom” is replaced by “to you”)

3.3 Morphological Analysis

To further preserve language semantics, we use morphological features. In this research, we pay attention to morphological features of Sinhala nouns only, since most of the rare words are noun word forms. We use two morphological features of Sinhala nouns,

1. Count (වචනය /*wachanaya*/)
2. Case (විභක්තිය /*wibhaktiya*/)

Grammatical number or the count is an inflectional feature belonging tothe realm of morphosyntax. Count can take three values, definite singular (DS), indefinite singular (IS) and definite plural (DP). Case is a suffix that is added to a stem to derive nouns in different meanings. Case forms differ in terms of the syntactic contexts in which they may occur. Sinhala language consists of 9 cases, ප්‍රථමා (/prathamaa/) - Nominative, කර්ම (/karma/) - Accusative, කර්තෘ (/kartru/) – Auxiliary, කරණ (/karaNa/) - Instrumental, සම්ප්‍රදාන (/sampradaana/) - Dative, අවදි (/awadi/) - Abalative, සම්බන්ධ (/sambandha/) - Possesive, ආධාර (/aadhaara/) - Locative, ආලපන (/aalapana/) - Vocative(Priyanga, Ranatunga& Dias, n.d.).

Synthetic parallel corpus that was generated in section 3.2 is further improved using morphological features. For a given word, there exists a variable number of case - count combinations. In this approach, we check whether the case - count combinations of the word that was removed have an intersection with the case - count combinations of the word that is introduced synthetically. We consider it as a syntax preserving sentence pair only if there exists an intersection of at least one element.

4. Experimental Setup

4.1 Data Collection and Preprocessing

Official government document translation is the domain used in this translation task. We used the parallel corpus developed by Farhath et al. (2017). Parallel corpus features government documents, annual reports, gazette papers, establishment codes, order papers, official letters and parliament documents. Characteristics of the Sinhala-Tamil parallel dataset are shown in Table 2.

Language	Total Words	Unique Words	Sentences
Sinhala	267,613	21,548	19,153
Tamil	226,160	38,651	

Table 2: Characteristics of the parallel dataset

Parallel corpus was divided into 3 parts: training set (14653 sentence pairs), validation set (4000 sentence pairs), and testing set (500 sentence pairs).

4.2 Experimental Setup

The open source NMT system: OpenNMT (Klein et al. 2017) was used for the experiments. GIZA++ (Och, Ney, 2004) was used for automatic word alignment. It uses the standard alignment heuristic grow-diag-final for word alignment. Tri-gram language models were trained for both source side and target side using the Stanford Research Institute Language Modeling toolkit (Stolcke et al, 2002) with Kneser- Ney smoothing. For translation evaluation, Bilingual Evaluation Understudy (BLEU) metric (Papineni et al, 2002) was used.

4.3 Benchmark Training

Using the above parallel corpus, Si-Ta and Ta-Si translation models were trained. Training involved two steps: pre-processing and model training. After completing the pre-processing step, two dictionaries (source dictionary and target dictionary) were generated to index mappings. Using two dictionaries and the serialized file, a model was trained with a 2-layer LSTM with 500 hidden units on both encoder and decoder.

4.4 Initial Data Augmentation

Out of 15383 number of unique words in the corpus, 6421 words appeared only once in the Sinhala language side, whereas corresponding values for Tamil side was found to be 31186 and 17238, respectively. Hence, we chose rare word threshold R to be 1. Considering the tradeoff between the number of sentences generated and semantic preservation of synthetic data, we chose fluency threshold M to be 2 and translation threshold T to be 0.9.

4.5 POS Tagging

Both original and synthetic parallel corpora that were generated in the previous section were POS tagged. We used the POS tagger developed by Fernando et al. (2016) for Sinhala, and the POS tagger developed by the Computational Linguistic Research Group (2017) for Tamil.

4.6 Morphological Analysis

We considered morphological features only when generating the Sinhala side of the parallel corpus. We used Helabasa - Noun Analyzer (2017) to retrieve Sinhala morphological features.

When training the translation model for each technique, we appended the synthetic corpus generated from that technique to our original corpus in one to one ratio and trained a separate model.

5. Results and Analysis

Table 3 provides examples resulting from each augmentation procedure.

Considering Table 3, in the initial data augmentation method (first row), substituting බඳවාගැනීමට (/banndhavaagAniemaTa/ - to hire) with දන්වන්නෙහිද (/danwannehida/ - will inform?) makes the resulting synthetic sentence meaningless. Sentences that are generated by analyzing POS tags seem to have an edge over initial data augmentation method. Since සැලස්මට (/salasmataTa/ - for plan) and බඳවාගැනීමට (/banndhavaagAniemaTa/ - to hire) (second row) have identical POS tags, high fluency is achieved in the resulting synthetic sentence pair. Synthetic sentence pair generated using the method mentioned in 3.3, preserves the meaning to a better extent. Word ලදුපතක්(/ladupatak/ - receipt) and කබායක්(/kabaayak/ - coat) have indefinite singular – Nominative, Accusative, Auxiliary, and Locative morphological features in common.

Method	Example
3.1	Si: එතුමා මෙම සභාවට [දන්වන්නෙහිද / **බඳවාගැනීමට**]? (/etumaa mema sabhaavaTa [danwannehida / **banndhavaagAniemaTa**] ?/) Ta: அவர் இச்சபைக்குத் [தெரிவிப்பாரா/ *ஆட்சேர்ப்பிற்கு*] ? (/avar issapaikkudh [dherivippaaraa/ *aatseerppiRku*] /) (En: For this session, he [will inform/ for hiring]?)[2]
3.2	Si: පුහුණු [සැලස්මට / **බඳවාගැනීමට**] අදාල තොරතුරු (/puhuNu [sAlAsmaTa / **banndhavaagAniemaTa**] adaala toraturu/) Ta: பயிற்சித் திட்டத்திற்கு [பொருத்தமான / *ஆட்சேர்ப்பிற்கு*] தகவல்கள் (/payiRsidh dhittadhdhiRku [porudhdhamaana / *aatseerppiRku*] dhakavalkaL/) (En: Information [related to training planning/ related to training hiring])[3]
3.3	Si: පහත සඳහන් ලිපිනයට කරුණාකර [ලදුපතක් / **කබායක්**] එවීමට කටයුතු කරන්න. (/pahata sanndhahan lipinayaTa karuNaakara [ladupatak / **kabaayak**]ewiemaTa kaTayutu karanna./) Ta: கீழ் காணும் முகவரிக்கு தயவு செய்து [பற்றுச் சீட்டொன்றை / *மழைக்காப்பு*]அனுப்ப நடவடிக்கை எடுக்கவும். (/kiiz kaaNum mukavarikku dhayavu seydhu [paRRus siittonRai / *mazaikkaappu*]anuppa watavatikkai etukkavum./) (En: Kindly send a [receipt/coat] for the following address)

Table 3: Examples synthetic data with highlighted [original / **substituted**] and [original /*translated*] words

Table 4 depicts the BLEU scores obtained for each method.

Method	Si-Ta	Ta-Si
Benchmark training	6.78	6.84
+ Initial data augmentation	7.68	8.86
+POS tagging	8.72	10.98
+Morphological Analysis	8.94	11.84

Table 4: BLEU scores

Synthetic data generated using the initial data augmentation method have improved the performance of Si-Ta and Ta-Si translation by 0.9 and 2.02 amounts, respectively. To verify that this gain is due to the rare word substitutions and not just due to the repetition of a part of the training data, we performed an experiment where each sentence pair selected for augmentation is added to the training data unchanged (i.e. without creating synthetic data). This simple form of sampled data replication delivered 0.53 and 1.42 BLEU score gains for Si-Ta and Ta-Si, respectively. Hence initial data augmentation models have performed better compared to simple data replication method.

Use of POS tags has achieved 1.04 and 2.12 BLEU score gains over the initial data augmentation for Si-Ta and Ta-Si, respectively. Human evaluators who oversaw the quality of generated sentences revealed that the use of POS tags has increased the fluency of language and rare word translation performance by a significant amount. Thus, we can empirically prove that the use of POS tags improves the quality of synthetic training data, which in turn reduces the rare word problem in NMT.

Morphological features have played a vital role in reducing the rare word problem. When generating synthetic sentence pairs, we considered only Sinhala language morphological features. Sinhala being morphologically rich, there exist many number of variations for a given root word. Hence checking the case-count combinations of a word when substituting, helps to preserve language semantics of the generated sentence. This is evident by analyzing the BLUE score gains of 1.26 and 2.98 for Si-Ta and Ta-Si translations compared to the initial data augmentation method.

Table 5 depicts an example of successful translation of a rare word.

Reference Source	விசேட பிரிவு(/*viseeta pirivu*/)
Reference Target	විශේෂ ඒකකය(/*viSeasha eakakaya*/) (Special unit)
Benchmark Model	විශේෂ අංශය(/*viSeasha a\nSaya*/) (Special sector) - erroneous
Our method (3.2 and 3.3)	විශේෂ ඒකකය(/*viSeashaeakakaya*/) (Special unit) - correct

Table 5: Rare Word example

To examine the impact of augmenting training data by creating contexts for rare words on the target side, we tested how each model performs on rare words. Most of the rare words are not 'rare' anymore in the augmented data since they were augmented sufficiently many times.

BLUE score gains are consistent across both translation directions, regardless of whether rare word substitutions are first applied to Sinhala or Tamil. Hence it can be verified that using POS tagging and morphological features results in generating quality synthetic parallel data that preserve language semantics, which eventually leads to better translation performance. Though overall rare word translation quality was improved by our methods, there were several cases where augmentation resulted in incorrect outputs that were correctly translated by our benchmark system. Table 6 corresponds to such an incorrect translation.

Our benchmark model has been able to correctly translate දේශීය (/*deaSieya*/ - local) and විදේශීය (/*videaSieya*/ - foreign) terms, whereas our new model has not been able to translate any of them. If the language model selects substitutions that have low probabilities, it results in generating outputs with low fluency. Another possible reason is errors in word alignments. If the word alignments are erroneous and phrase table contains faulty

[2] Exact English translation:"Will he inform this session?" / "For this session will he be hiring?" ("will inform" is replaced by "for hiring")

[3] Exact English translation:"Information related for training planning" / "Information related for training hiring"("for training" is replaced by "for hiring")

probabilities, this may lead to synthetic sentence pairs that do not correspond to each other.

Source Sentence	8. உள்ளூர்/ வெளிநாட்டு திரைப்ப ட தயாரிப்பாளர்களுடன் /uLLur/ veLiwaattu dhiraippa ta dhayaarippaaLarkaLutan/ - (With local and foreign film producers)
Reference Translation	8 .දේශීය / විදේශීය චිත්‍රපට නිෂ්පාදකයින් සමඟ /8 .deaSieya / videaSieya citrpaTa nishpaadakayin samannga / - (With local /foreign film producers)
Benchmark translation	8 .දේශීය විදේශීය විකාශන කටයුතු සඳහා/ 8 .deaSieya videaSieya vikaaSana kaTayutu sanndhahaa/ - (For local/ foreign broadcasting)
Our method (3.2 and 3.3)	(8)විදේශ චිත්‍රපට ගනුදෙනු කිරීම / (8)videaSa citrapaTa ganudenu kiriema . / - (For trading foreign films)

Table 6: Incorrect outputs

6. Conclusion

The purpose of this research was to find out syntax preserving techniques for synthetic data generation to solve the rare word problem in NMT for the under-resourced language pair Sinhala and Tamil. POS tagging and morphological analysis show impressive results in increasing the quality of synthetic sentence pairs that reduces the rare word problem. Being morphologically rich, there exist a number of morphological features in Sinhala and Tamil that can be exploited to enhance the quality of augmented data. We expect to experiment with these features in the future.

7. Acknowledgements

The authors would like to thank the reviewers for their helpful comments and suggestions. The authors are grateful to members of the National Languages Processing Centre at University of Moratuwa for their significant contribution in developing the basic linguistic resources, and the Department of official languages of Sri Lanka for providing corpus data needed to carry out the research.

8. Bibliographical References

Bahdanau, D., Cho, K., &Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. Arxiv Preprint Arxiv:1409.0473 [Cs.CL].

Cho, K., van Merrienboer, B., Bahdanau, D., &Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8) ,pp. 103-111.

Fadaee, M., Bisazza, A., &Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics ,pp. 567-573.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & M. Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation",.Arxiv Preprint Arxiv:1701.02810 [Cs.CL],.

Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. Computational linguistics, 30(4),pp. 417-449.

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

Priyanga, R., Ranatunga, S., & Dias, G. An Inflectional Morphological Generator for Sinhala Nouns. (Unpublished)

Pushpananda, R., Weerasinghe, R., & Niranjan, M. (2014). Sinhala-Tamil Machine Translation: Towards better Translation Quality. In proceedings of Australasian Language Technology Association Workshop, 129, pp. 129-133.

Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh Neural Machine Translation Systems for WMT 16. In Proceedings of the First Conference on Machine Translation (WMT) ,pp. 371-376. Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 86-96.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In Interspeech. pp. 901-904.

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., Ranathunga, S., Dias, G., &Jayasena, (2017). Neural Machine Translation for Sinhala and Tamil Languages, *(Submitted).*

9. Language Resource References

Computational Linguistic Research Group. (2017). Au-kbc.org. Retrieved 2 October 2017, from http://www.au-kbc.org/nlp/corpusrelease.html

Farhath, F., Ranathunga, S., Jayasena, S., Dias, G., &Thayasivam, U. (2017). Improving Domain-Specific Statistical Machine Translation for Sinhala-Tamil using Bilingual Lists, *(Submitted).*

Fernando, S., Ranathunga, S., Jayasena, S., & Dias, G. (2016). Comprehensive Part-Of-Speech Tag Set and SVM based POS Tagger for Sinhala. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (pp. 173 - 182).

Helabasa - Noun Analyzer. (2017). Translation.projects.mrt.ac.lk. Retrieved 1 October 2017, from http://translation.projects.mrt.ac.lk:8081/helabasa/noun_analyzer