Neural machine translation for sinhala and tamil languages

Conference Paper · December 2017

DOI: 10.1109/IALP.2017.8300576

7 authors, including:

Pasindu Tennage
École Polytechnique Fédérale de Lausanne
10 PUBLICATIONS **63** CITATIONS
SEE PROFILE

Prabath Sandaruwan
University of Moratuwa Sri Lanka
2 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Malith Thilakarathne
University of Moratuwa
3 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Achini Herath
University of Moratuwa
3 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Neural Machine Translation for Sinhala and Tamil Languages

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

***Abstract*—Neural Machine Translation (NMT) is becoming the current state of the art machine translation technique. Although NMT is successful for resourceful languages, its applicability in low-resource settings is still debatable. In this paper, we address the task of developing a NMT system for the most widely used language pair in Sri Lanka- Sinhala and Tamil, focusing on the domain of official government documents. We explore the ways of improving NMT using word phrases in a situation where the size of the parallel corpus is considerably small, and empirically show that the resulting models improve our benchmark domain specific Sinhala to Tamil and Tamil to Sinhala translation models by 0.68 and 5.4 BLEU, respectively. The paper also presents an analysis on how NMT performance varies with the amount of word phrases, in order to investigate the effects of word phrases in domain specific NMT.**

Keywords- Neural Machine Translation (NMT); Word Phrases

I. INTRODUCTION

Neural Machine Translation (NMT) is a new architecture that aims at building a single neural network that can be jointly tuned to maximize translation performance. NMT delivers state of the art results, especially for language pairs involving rich morphology prediction and significant word reordering. NMT generates outputs that have lower post-edit effort with respect to Statistical Machine Translation (SMT) outputs. NMT seems to have an edge especially on lexically rich texts [1]. However, performance of NMT degrades rapidly as the parallel corpus size gets small.

Sinhala and Tamil languages are under resourced due to lack of sufficiently large parallel corpora. Existing translation models for this language pair have not yet reached the stage of proliferation demonstrated by translation systems for other languages [2]. Thus, the use of NMT for Sinhala and Tamil is challenging.

In this research, we developed a domain specific NMT system for language pair - Sinhala and Tamil, the official languages in Sri Lanka. Official government document translation was focused in this research. In this research, we used the NMT architecture proposed by Bahdanau et al. [3] and Cho et al. [4] for all the experiments.

Given that word phrases play an important role in domain specific machine translation, we explore effective methods of using word phrases to improve NMT performance for these two under resourced languages. We discuss how NMT performance varies with the amount of word phrases added and the effect of word phrases in domain specific NMT. We also use an existing method of using monolingual target side data to improve domain specific NMT performance [5].

Finally, the paper highlights the effect of sentence length for translation performance for Sinhala and Tamil NMT.

II. BACKGROUND AND RELATED WORK

A. Neural Machine Translation

NMT is an end-to-end translation process [3], which does not rely on pre-designed feature functions. The goal of NMT is to design a model of which every component is tuned based on training corpora to maximize its translation performance. Encoder Decoder architecture with attention mechanism is the current state of the art NMT architecture.

Recurrent activation function is applied recursively over the input sentence, until the end when the h_T which is the final internal state of the recurrent neural network (RNN) contains the summary of the whole input sequence. After the last word's continuous vector s_T (T is input sequence length) is read, the RNN's internal state h_T represents a summary of the whole source sequence. Decoder computes RNN's internal state z_i based on the summary vector h_T, the previous predicted word u_{i-1} and the previous internal state z_{i-1}. Using decoder's internal hidden state z_i, it's possible to score each target word based on how likely it is to follow all the preceding translated words. Once the score of every word is computed, using softmax normalization, scores are turned into proper probabilities.

There have been multiple efforts to improve performance of NMT. In recent research, data augmentation to increase NMT performance for under resourced languages has been explored [6]. In this work, synthetic parallel sentences have been generated to increase the number of rare word occurrences. Using target side monolingual data to improve NMT performance for general under resourced languages has also been proposed [5]. According to this research, using synthetic source side sentences generated from back translation has increased the quality of translation by a significant amount.

A method to translate phrases in NMT by integrating a phrase memory into the encoder-decoder architecture of NMT has been presented by Wang et al. [7]. In this model, at each decoding step, the phrase memory is first rewritten by the SMT model, which dynamically generates relevant target phrases with contextual information provided by the NMT model.

B. Sinhala - Tamil machine translation

Sinhala language descends from Indic language family and Tamil from Dravidian family [8]. Being morphologically rich, Sinhala has up to 110 noun word forms and up to 282 verb word forms [9] and Tamil has

978-1-5386-1981-0/17/$31.00 ©2017 IEEE

around 40 noun word forms and up to 240 verb word forms [10]. Both these languages have the same word order of Subject-Object-Verb. However, both languages have the flexibility to alter the word order.

There is no published literature on applying NMT for Sinhala and Tamil machine translation. However, some research has been carried out on Sinhala-Tamil SMT [11]. Sinhala-Tamil language pair gives better performance compared to the Sinhala-English pair in SMT due to similarities between Sinhala and Tamil [11].

Recently a research has been carried out on development of a Sinhala-to-Tamil SMT system for official government documents "unpublished" [12]. This system has been developed with emphasis given to domain adaptation. Performance of the system has been evaluated with the static integration of three types of lists, namely, a list of government organizations and official designations, a glossary related to government administrations and operations, and a general bilingual dictionary to the translation model of the SMT system.

III. Methodology

In this research, we identified a novel approach of using word phrases to improve domain specific NMT performance. We also empirically tested the applicability of using target side monolingual training data to improve the performance of Sinhala to Tamil NMT, as done by Sennrich et al. [5].

A. Including Word phrases

We consider a word phrase as a combination of 1 or more words that has a specific meaning when taken together. Maximum word phrase size was set to 3 words. Several types of word phrases that are in domain with this translation task were extracted and added to the training corpus. These include a set of named entities, a set of common domain specific terms and phrases, a set of government designations and frequently used phrases that are used in government documents. These word phrases were integrated statically into the NMT system.

B. Monolingual Training Data

Monolingual data are especially helpful if parallel data are sparse, or if there is a poor fit for the translation task, for instance because of a domain mismatch. Techniques that can be used to improve the quality of NMT using monolingual data have been identified by Sennrich et al. [5]. According to the authors, adding target side monolingual data where the source side data are generated using automatic back translation produces better results compared to using dummy source sentences. Hence, we used the synthetic source sentence method to increase the performance of NMT. Back translation of target side monolingual data was done using our system itself.

IV. Experimental Setup

A. Data

The domain of this translation task is official government documents of Sri Lanka. We used the parallel corpus developed by Farhath et al. "unpublished" [12]. Parallel corpus features government documents such as annual reports, establishment codes, order papers, and official letters. The extracted parallel data have been manually cleaned with the help of a custom developed tool. Human translators oversaw the process of extracting data from the above-mentioned documents and ensured the validity of the translation materials used in the data set. Statistics of the Sinhala-Tamil parallel dataset are shown in Table I.

Parallel corpus was divided into 3 parts: training set, validation set, and testing set. Each dataset consisted of parallel source and target data containing one sentence per line with tokens separated by a space. Validation files were used to evaluate the convergence of the training. To make an unbiased test data set, it was necessary to take the relevant ratios of sentence pairs from different sources.

B. Pre-Processing

Both Sinhala and Tamil languages contain one or more symbols per character, unlike English. Due to this characteristic of Sinhala and Tamil, existing tokenization tools were not able to tokenize the text, since they identified a single character as two characters. Hence a tokenizer that was specifically developed for Sinhala and Tamil was used in this research.

C. System Setup

The open source NMT system OpenNMT [13] was used for the experiments. OpenNMT supports standard encoder - decoder architecture with attention mechanism.

To evaluate the quality of the translation, Bilingual Evaluation Understudy (BLEU) metric [14] was used. We used the percentage BLEU score values in this paper (0 to 100 range).

D. Benchmark Training

Using the above parallel corpus, two benchmark systems were trained: Sinhala to Tamil, and Tamil to Sinhala. Training involved two different 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 these two dictionaries and the serialized file, a model was trained with 2-layer Long Short-Term Memory with 500 hidden units on both encoder and decoder. Since most of the operations inside the network were numeric and easily parallelizable, NVIDIA TESLA C2070 with GPU memory 5.5 GB was used to speed up the process.

E. Including Word Phrases

Four types of word phrases that are in domain with this translation task were extracted and added to the training corpus.

TABLE I. Characteristics of the Parallel Dataset

Language	Total Words	Unique Words	Sentences
Sinhala	346030	19531	23611
Tamil	293821	37243	

- Set of named entities-11,561 pairs
- Set of common domain specific terms and phrases- 19,861 pairs
- Set of government designations- 5,291 pairs
- Frequently used phrases that are used in government documents (Letter heads, salutations etc.) - 610 pairs

To find the effect of number of word phrases for the BLEU score, a comprehensive analysis was carried out. We trained separate models for Sinhala to Tamil, and Tamil to Sinhala by adding 5000 more-word phrases to the initial training dataset each time. Experiments were carried out for 5k, 10k, 15k, 20k, 25k, 30k, 35k, 40k, 45k and 47k number of word phrases.

F. Monolingual Training Data

Target side in-domain monolingual data were extracted using official letters. Maximum sentence length was set to 30 to avoid the performance issues in NMT systems when sentence length is large. 10, 000 target side parallel sentences for each language were extracted and back translated. Model that was trained using the extended corpus that included word phrases was used to back translate. Since the generated synthetic source side data had translation errors, we restricted number of synthetic sentence pairs to be less than the number of sentence pairs in the original parallel corpus, to make sure that the overall quality of resulting parallel corpus remains acceptable. Two models were trained for Sinhala to Tamil and Tamil to Sinhala separately.

Another analysis was carried out to identify how BLEU score is affected by sentence length for Sinhala Tamil NMT. In this study, we calculated average BLEU score for each group of sentences with a particular length.

V. RESULTS AND ANALYSIS

Table II depicts the BLEU scores obtained for each method.

Compared to results achieved in general machine translation tasks [2, 8], the results we achieved were significantly higher with respect to the dataset size we used. Major reason for this mismatch is the domain-specific nature of the dataset we used. Since the vocabulary and language constructs are smaller in our dataset, compared to datasets in general machine translation tasks, the model is fine tuned for the domain. Hence the BLEU scores are high.

The results obtained were contradicting with the results obtained by Bentivogli et al. [1], which states that NMT performs better than SMT for the same corpus size. For the small dataset we have, SMT performed better than NMT, according to our observations. Number of parameters that need to be learnt in the training process is higher in NMT compared to SMT, which leads to this observation.

Adding word phrases has increased the BLEU score by 0.68 for Sinhala to Tamil translation, and by 5.4 for Tamil to Sinhala translation. This BLEU score gain is due to two main factors. Firstly, adding word phrases increases the corpus size. It should be noted that the word phrases that were added are not complete sentences, but contain only 2-3 words per phrase.

Second reason for the BLEU score gain is the nature of domain-specific language translation behavior. In this research, we used official government documents as our domain. The word phrases included a significant amount of named entities that are widely used in official government documents. Even though there is no explicit language model in NMT, the decoder considers the last translated word when assigning the probability to the next translated word. There is a high chance that a named entity appears only once in the original training corpus. Hence, a low probability would be applied for the correct next word, due to low presence of two adjacent words in the corpus. Adding word phrases helps to increase this probability, thus reducing rare word problem.

Fig. 1 shows the graph that depicts BLEU score against the number of word phrases.

When the number of word phrases is increased by 5000, the increase in BLEU score is 0.0723 BLEU points in average for Sinhala to Tamil translation and 0.5744 BLEU points for Tamil to Sinhala translation.

Use of target side monolingual data improved the translation quality by 0.13 for Sinhala to Tamil and 3.43 for Tamil to Sinhala. A central theoretical expectation is that monolingual target-side data improve the model's fluency, and its ability to produce natural target-language sentences. This BLEU score gain is due to two factors. Target side monolingual data play a vital role in language modeling in SMT. Being an end to end process, NMT does not have a separate language model. Yet in the decoder, NMT system considers the previously translated word when predicting the new translation. Hence when in-domain target side monolingual data are added to the training corpus by automatically back translating them to source side, NMT system can take advantage of language specific features in the target side. Second major reason for BLEU score gain is the increased corpus size. Even though the quality of back translated source side is low,

TABLE II. BLEU SCORES

Method	Sinhala-Tamil	Tamil-Sinhala
Benchmark system	6.78	6.84
+Adding word phrases	7.46	12.24
+Monolingual training data	6.91	10.27
+Adding word phrases +Monolingual training data	7.50	12.75
SMT [12]	17.06	Not Trained

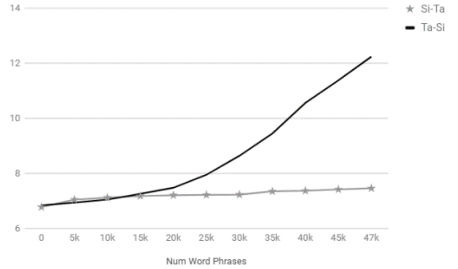

Figure 1. Number of word phrases vs BLEU score

compared to our original parallel corpus, the model output was increased due to the increased number of sentences in the parallel corpus.

According to Table II, Sinhala to Tamil BLEU score is worse than Tamil to Sinhala BLEU score for every method. In Table I, for the same parallel corpus, number of words and unique words in Tamil are greater than respective values for Sinhala. Hence when translating from Sinhala to Tamil, out of vocabulary problem is more significant, compared to Tamil to Sinhala.

Fig. 2 depicts the relationship between sentence length and BLEU score for both Sinhala to Tamil, and Tamil to Sinhala translations, respectively.

Both models have performed better with shorter sentences than longer sentences. As the sentence length increases, the BLEU score has dramatically decreased for both models.

VI. CONCLUSION

The purpose of this research was to improve performance of NMT when corpus size is small. We can conclude that while Tamil to Sinhala and Sinhala to Tamil translations are unable to produce intelligible output with a parallel corpus of just 23611 sentence pairs, we can improve the translation performance by adding word phrases and using monolingual training data. We can expect performance to approach usable levels by collecting a large parallel corpus. Using this experience, we are currently collecting a more balanced parallel corpus.

Morphological richness in the two languages is one of the major reasons to get lower results. Furthermore, a preliminary study shows that it is possible to improve performance for the same dataset we used for this research by treating words at the character level rather than word level [15]. In future, we are planning to investigate on the applicability of this character level NMT approach for Sinhala and Tamil. We will continue to improve this NMT system to a level that it is capable of producing acceptable translations between Sinhala and Tamil for use by the wider community.

This work paves way for new research topics related to word phrases. Applicability of word phrases for big data applications and domain adaptation are possible future work.

ACKNOWLEDGMENT

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.

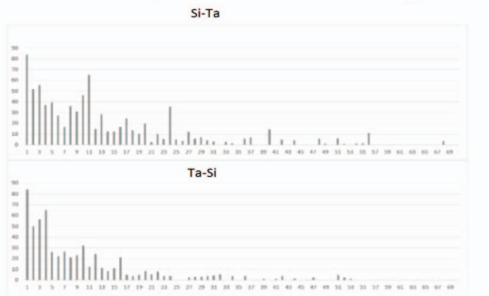

Figure 2. Sentence length vs BLEU score

REFERENCES

[1] L. Bentivogli, A. Bisazza, M. Cettolo and M. Federico, "Neural versus Phrase-Based Machine Translation Quality: a Case Study", in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA, 2016, pp 257-267.

[2] R. Sennrich, B. Haddow and A. Birch, "Edinburgh Neural Machine Translation Systems for WMT 16", in *Proceedings of the First Conference on Machine Translation (WMT)*, Association for Computational Linguistics, 2016, pp. 371-376.

[3] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *arxiv preprint arXiv:1409.0473 [cs.CL]*, 2014.

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

[5] R. Sennrich, B. Haddow and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data", in *Proceedings of the Annual Meeting of the Association for Computational Linguistics,*ACL, 2016,pp. 86-96.

[6] M. Fadaee, A. Bisazza and C. Monz, "Data Augmentation for Low-Resource Neural Machine Translation.", in *Proceedings of the Annual Meeting of the Association for Computational Linguistics* ,ACL , 2017, pp. 567-573.

[7] X. Wang, Z. Tu, D. Xiong and M. Zhang, "Translating Phrases in Neural Machine Translation", in *Conference on Empirical Methods in Natural Language Processing*, 2017,pp. 1432-1442..

[8] P. Randil, R. Weerasinghe and M. Niranjan, "Sinhala-Tamil Machine Translation: Towards better Translation Quality", *Australasian Language Technology Association Workshop 2014*, vol. 129, pp. 129-133.

[9] V. Welgama, D. Herath, C. Liyanage, N. Udalamatta, R. Weerasinghe and T. Jayawardana, "Towards a Sinhala Wordnet", in *Conference on Human Language Technology for Development*, Alexandria, Egypt, 2011, pp. 39-43.

[10] S. Lushanthan, A. Weerasinghe and D. Herath, "Morphological analyzer and generator for tamil language", in *IEEE conference on Advances in ICT for Emerging Regions (ICTer),* , 2014, pp. 190-196.

[11] R. Weerasinghe, "A statistical machine translation approach to sinhala-tamil language translation.", *Towards an ICT enabled Society*, 2003 pp. 136-141.

[12] F. Farhath, "Sinhala-to-Tamil Machine Translation of Short Official Documents".

[13] G. Klein, Y. Kim, Y. Deng, J. Senellart and A. M. Rush, "OpenNMT: Open-source toolkit for neural machine translation", *arxiv preprint arXiv:1701.02810 [cs.CL]*, 2017.

[14] K. Papineni, S. Roukos, T. Ward and W. Zhu, "BLEU: a method for automatic evaluation of machine translation.", in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002, pp. 311-318.

[15] Chung, K. Cho and Y. Bengio, "A Character-Level Decoder without Explicit Segmentation for Neural Machine Translation", in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*,ACL 2016, pp. 1693-17