Sinhala-Tamil Machine Translation: Towards better Translation Quality

Randil Pushpananda **Ruvan Weerasinghe**
Language Technology Research Laboratory
University of Colombo School of Computing
Sri Lanka
{rpn|arw}@ucsc.lk

Mahesan Niranjan
School of Electronics and
Computer Science
University of Southampton, UK
mn@ecs.soton.ac.uk

Abstract

Statistical Machine Translation (SMT) is a well-known and well established data-driven approach used for language translation. The focus of this work is to develop a statistical machine translation system for Sri Lankan languages, Sinhala and Tamil language pair. This paper presents a systematic investigation of how Sinhala-Tamil SMT performance varies with the amount of parallel training data used, in order to find out the minimum needed to develop a machine translation system with acceptable performance.

1 Introduction

Sri Lanka is a multi-ethnic, multi-lingual country. Sinhala and Tamil are the national languages of Sri Lanka. The majority of Sri Lankans do not have a good knowledge of languages other than their mother tongue. Therefore a language barrier between the Sinhala and Tamil communities exists. This language barrier and the problems that arose during the last 30 years in the country, encouraged us to a translation application using the SMT approach. This would reduce the language gap between these two communities and thereby help solve a burning issue in the country.

The choice of the Sinhala - Tamil language pair provides some opportunities as well as some challenges. The opportunity is that they share some affinity to each other, having evolved alongside each other in Sri Lanka. The challenges include the sparseness in the availability of data, and the limited research undertaken in them. Hence, developing a successful system with limited resources is our ultimate goal.

2 Background and Related Work

There is very limited research reported in the literature for Sinhala-Tamil machine translation. According to (Weerasinghe, 2003), the Sinhala-Tamil language pair gives better performance compared to the Sinhala-English pair in SMT since they are more closely related to each other owing to their evolution within Sri Lanka. Some important factors to consider when building SMT for the Sinhala-Tamil language pair have been identified in (Sakthithasan et al., 2010). The limited amount of data, and the restricted domain it represented, makes that word hard to generalize. Another study (Jeyakaran and Weerasinghe, 2011), explored the applicability of the Kernel Ridge Regression technique to Sinhala-Tamil translation. This research resulted in a hybrid of classical phrase based SMT and Kernel Ridge Regression with two novel solutions for the pre-image problem.

Owing to the limited amount of parallel data available, it has been not possible to analyze how the results vary with increasing numbers of parallel sentences in Sinhala and Tamil for general purpose MT.

2.1 Sinhala and Tamil Languages

Sinhala belongs to the Indo-Aryan language family and Tamil to the Dravidian family. Both Sinhala and Tamil languages are morphologically rich languages: Sinhala has up to 110 noun word forms and up to 282 verb word forms (Welgama et al., 2011) and Tamil has around 40 noun word forms and up to 240 verb word forms (Lushanthan, 2010). Also both these languages are syntactically similar. The typical word order of both these languages are Subject-Object-Verb. However both are flexible with the word order and variant word orders are possible with discourse - pragmatic effects (Liyanage et al., 2012; Wikipedia, 2014).

In addition there are some of the aspects of Tamil influence on the structure of the Sinhalese language. The most significant impact of Tamil on Sinhalese has been at the lexical level (Karunatilaka, 2011). අම්මා (/amma/: mother), අක්කා

Randil Pushpananda, Ruvan Weerasinghe and Mahesan Niranjan. 2014. Sinhala-Tamil Machine Translation: Towards better Translation Quality. In *Proceedings of Australasian Language Technology Association Workshop*, pages 129−133.

(/akka/: elder sister), අයියා (/ayya/: elder brother) are some loan words out of more than thousand words borrowed from Tamil to Sinhala (Coperahewa and Arunachalam, 2011).

3 Experiments and Results

3.1 Tools used

The open source statistical machine translation system: MOSES (Koehn et al., 2007) was used with GIZA++ (Och and Ney, 2004) using the standard alignment heuristic grow-diag-final for word alignments. Tri-gram language models were trained on the target side of the parallel data and the target language monolingual corpus by using the Stanford Research Institute language Modeling toolkit (Stolcke and others, 2002) with Kneser-Ney smoothing. The systems were tuned using a small extracted parallel dataset with Minimum Error Rate Training (MERT)(Och, 2003) and then tested with different test sets. Finally, the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) evaluation metric was used to evaluate the output produced by the translation system.

3.2 Data Collection and Data Preprocessing

To build a good baseline system, we need to have a sentence-aligned parallel corpus to train the translation model and a (possibly larger) monolingual corpus of the target language to train the language model.

Language	**Characteristics**		
	Total Words	**Unique Words**	**Sentences**
Sinhala	10,142,501	448,651	850,000
Tamil	4,288,349	400,293	407,578

Table 1: Characteristics of Sinhala and Tamil Monolingual Corpora

We used the *UCSC*[1] *10M words Sinhala Corpus* (Weerasinghe et al., 2007) and the *4M words Tamil Corpus* (Weerasinghe et al., 2013) to build the Sinhala and Tamil language models respectively. Both these are open domain corpora mainly with newspaper articles and Technical writing. The characteristics of the Sinhala and Tamil corpora is shown in Table 1.

Finding a good large Sinhala-Tamil parallel corpus was the main difficulty. For this purpose we collected a *Sinhala-Tamil Parallel Corpus* (Weerasinghe and Pushpananda, 2013) which consists of 25500 parallel sentences. This is also an open domain corpus which includes mainly newspaper texts and technical writing. The sentence length of sentences in this corpus was restricted to 8 - 12 words. Both Sinhala to Tamil and Tamil to Sinhala translation models were built using this corpus. The characteristics of the Sinhala-Tamil parallel dataset is shown in Table 2

Language	**Total Words(TW)**	**Unique Words(UW)**	**UW/TW**
Sinhala	252,101	37,128	15%
Tamil	219,017	53,024	24%

Table 2: Characteristics of parallel dataset

3.2.1 Baseline Systems

Using the above parallel corpus, we trained two baseline systems: Sinhala to Tamil and Tamil to Sinhala. First, 500 parallel sentences were extracted randomly as the tuning dataset. Then of the remaining 25000 parallel sentences, 5000 sentences were extracted randomly as the initial dataset. By applying *10-fold cross-validation* (Kohavi and others, 1995) (to get an unbiased result), we divided extracted 5000 sentences into 10 mutually exclusive partitions equally and then one of the partitions was used as the testing data and the other nine used as training data. Then we trained and evaluated the system iteratively for all combinations of the datasets and finally calculated the average performance of the results in order to obtain unbiased estimates of accuracy. We repeated the same procedure by adding 5000 more sentences to the initial dataset each time until the remaining dataset was empty.

Results Figure 1 shows the average BLEU score value variation against the number of parallel sentences in both Sinhala to Tamil and Tamil to Sinhala translation. However, it clearly indicates that much more data would be required to build an acceptable translation model for the Sinhala-Tamil language pair. The results of the Tamil to Sinhala translation system in figure 1 shows that the BLEU score approaches 12.9 when the dataset size reaches 25000. It also shows that results of the Sinhala to Tamil translation only approaches 10.1 for the full dataset of 25000 parallel sentences. The figure 1 shows that when the dataset size is increased from 5000 to 10,000 and 10,000 to 20,000, the increase in performance varies by

[1] University of Colombo School of Computing

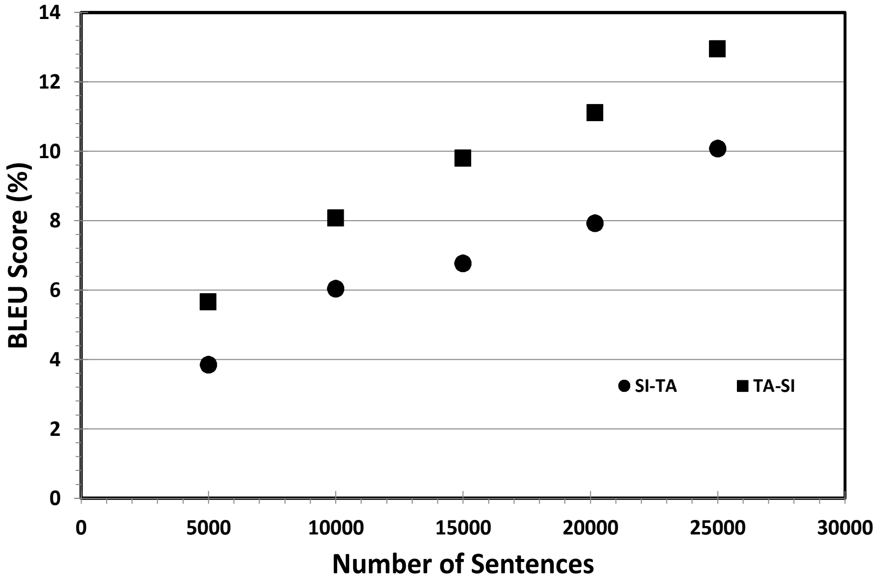

Figure 1: Average BLEU Score VS Number of Parallel Sentences

around 2 BLEU points for Sinhala to Tamil translation and around 2 to 3 BLEU points for Tamil to Sinhala translation. This is consistent with the results reported by Turchi et al. (2012).

Language	Sample Size (S)	Average Perplexity	Out of Vocabulary (OOV)	OOV/S
Sinhala	5000 25000	1590.10 997.33	962 2225	19% 9%
Tamil	5000 25000	6067.65 3819.94	1295 3593	26% 14%

Table 3: Average perplexity values and out-of-vocabulary values of the Sinhala-Tamil Parallel Corpus

Also, as shown in table 3, we can clearly see that as the number of sentences are increased, the average perplexity for both Sinhala and Tamil decreases. Sinhala and Tamil datasets were considered separately from the parallel corpus to calculate the perplexity values. These values are very high compared to those of the dominant European languages.

Here we did an error analysis to identify the problems of the methods we used and to find new methodologies to improve the results.

4 Error Analysis

The BLEU scores for test sets of 5000 and 25000 data samples were taken for the error analysis. The process for the error analysis stated as follows.

- Calculate the number of total words(TotW) and unique words(UniW) in each training (Tr) and test (Te) datasets.
- Calculate the number of out-of-vocabulary (OOV) words in the test dataset (as a percentage of test dataset).
- Calculate the number of untranslated words (UntransW)(as a percentage of test dataset).
- Calculate the number of translated words which are not in the reference dataset (TargetOOV)(as a percentage of test dataset).
- Calculate the number of translated words which are not in the target language model (Target LM OOV) (as a percentage of reference dataset).

Description	5000		25000	
	TotW	UniW	TotW	UniW
Training Dataset	44,806	13,723	224,959	34,858
Testing Dataset	4,985	2,884	24,678	8,890
OOV (%)	19.70	33.29	9.41	25.11
UntransW (%)	33.78	52.98	17.82	44.26
Reference Dataset	3,168	1,307	17,584	4,298
TargetOOV (%)	17.65	19.15	9.58	17.43
Target LM OOV (%)	0.29	0.33	1	1.55

Table 4: Results obtained from the error analysis of Sinhala to Tamil translation

The results obtained for the Sinhala to Tamil and Tamil to Sinhala translations are shown in

Description	5000		25000	
	TotW	UniW	TotW	UniW
Training Dataset	39,044	16,328	194,784	49,402
Testing Dataset	4,336	2,968	21,462	10,381
OOV (%)	30.32	43.67	16.84	33.85
UntransW (%)	40.68	57.14	25.08	48.58
Reference Dataset	3,168	1,307	17,584	4,298
TargetOOV (%)	10.88	14.94	5.01	11.45
Target LM OOV (%)	0.04	0.07	0.15	0.44

Table 5: Results obtained from the error analysis of Tamil to Sinhala translation

table 4 and 5 respectively. When considering the 5000 and 25000 datasets in table 4 and 5, we can see that the total number of words in the Tamil to Sinhala translation is lower than the Sinhala to Tamil translation in both training and testing datasets. However the unique number of words in the Tamil to Sinhala translation is much higher than the Sinhala to Tamil translation. This clearly shows the complexity of the Tamil language. However, as we expected OOV (unique word) rate is reduced by 8% - 10%, when the dataset size is increasing. That is one of the reasons for the increment of BLEU score value. We have identified mainly two problems. According to table 4, 20% of unique words in the test set are not translated even they were in the training set and 17% to 19% of words which are not in the target reference set is in the translated output. Those are occurred due to phrase alignment problems and also the decoding problems. For an example if we need to translate ගෙදර (*Home*) to Tamil, the phrase table consists only ගෙදර එන්න (*Come home*) and ගෙදර යන්න (*Go home*), then that word will not be translated even that word is in the training set. Since Sinhala and Tamil are low-resourced languages, we need to consider these issues to build a good translation system. We can clearly see that out-of-vocabulary rate and the untranslated word rate is much higher in Tamil to Sinhala Translation. Also when we consider the out-of-vocabulary words, we have found that those words consist of proper names, misspelled words, inflections, derivatives and honorifics. These are the main problems that we could identify from the error analysis. Since human evaluation is very costly, we used only the above technique to do the evaluation. According to the figure 1, we can see that even the OOV words are higher, BLEU score values of Tamil to Sinhala translation is higher. The main reason for this could be the size of the language model since words in the Sinhala monolingual corpus is more than twice as the words in the Tamil monolingual corpus. When consider the Target OOV and Target LM OOV in Tamil to Sinhala Translation is lower compared to the Sinhala to Tamil translation. That could be a another reason to get a higher BLEU score value for Tamil to Sinhala translation.

5 Conclusion and Future Work

The purpose of this research was to find out how the SMT systems perform for Sinhala to Tamil and Tamil to Sinhala translation. We can conclude that while Tamil to Sinhala and Sinhala to Tamil translation is unable to produce intelligible output with parallel corpus of just 25000 sentence pairs of relatively short length, we can expect performance to approach usable levels by collecting a large parallel corpora. Using this experience, we are currently collecting a more balanced parallel corpus.

However the error analysis shows that the sentence length limitations of the Sinhala-Tamil parallel corpus could not be the only reason for the comparatively lower BLEU scores, morphological richness may be the reason to get lower results since misspelled words and proper names are common to other languages too. Furthermore, a preliminary study shows that we can get better perplexity values for the same dataset we used for this research by stemming suffixes of the Sinhala and Tamil parallel sentences. In future, we are planning to investigate and find solutions to these problems and planning to implement a system capable of producing acceptable translations between Sinhala and Tamil for use by the wider community.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Research Council, ICT Agency and LK Domain Registry of Sri Lanka. The authors are grateful to past and current members of the Language Technology Research Laboratory of the UCSC, Sri Lanka for their significant contribution in developing the basic linguistic resources needed to carry out the research described above.

References

Sandagomi Coperahewa and Sarojini Arunachalam. 2011. *A Dictionary of Tamil Word in Sinhala*, volume 2. Godage International publishers, Sri Lanka.

Mahendran Jeyakaran and Ruvan Weerasinghe. 2011. A novel kernel regression based machine translation system for sinhala-tamil translation. In *Proceedings of the 4th Annual UCSC Research Symposium*.

WS Karunatilaka. 2011. *Link*. Godage International publishers, Sri Lanka.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, 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*, pages 177–180. Association for Computational Linguistics.

Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145.

Chamila Liyanage, Randil Pushpananda, Dulip Lakmal Herath, and Ruvan Weerasinghe. 2012. A computational grammar of sinhala. In *Computational Linguistics and Intelligent Text Processing*, pages 188–200. Springer.

Sivaneasharajah Lushanthan. 2010. Morphological analyzer and generator for tamil language. August.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Sripirakas Sakthithasan, ruvan Weerasinghe, and Dulip Lakmal Herath. 2010. Statistical machine translation of systems for sinhala-tamil. In *Advances in ICT for Emerging Regions (ICTer), 2010 International Conference on*, pages 62–68. IEEE.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

Marco Turchi, Cyril Goutte, and Nello Cristianini. 2012. Learning machine translation from in-domain and out-of-domain data. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 305–312.

Ruvan Weerasinghe and Randil Pushpananda. 2013. Sinhala tamil parallel corpora subset with a total 1 million words. Technical report, University of Colombo School of Computing.

Ruvan Weerasinghe, Dulip Herath, Viraj Welgama, Nishantha Medagoda, Asanka Wasala, and Eranga Jayalatharachchi. 2007. Ucsc sinhala corpus - pan localization project-phase i.

Ruvan Weerasinghe, Randil Pushpananda, and Namal Udalamatta. 2013. Sri lankan tamil corpus. Technical report, University of Colombo School of Computing and funded by ICT Agency, Sri Lanka.

Ruvan Weerasinghe. 2003. A statistical machine translation approach to sinhala-tamil language translation. *Towards an ICT enabled Society*, page 136.

Viraj Welgama, Dulip Lakmal Herath, Chamila Liyanage, Namal Udalamatta, Ruvan Weerasinghe, and Tissa Jayawardana. 2011. Towards a sinhala wordnet. In *Proceedings of the Conference on Human Language Technology for Development*.

Wikipedia. 2014. Tamil language — wikipedia, the free encyclopedia. [Online; accessed 30-October-2014].