Statistical Machine Translation from and into Morphologically Rich and Low Resourced Languages

Conference Paper *in* Lecture Notes in Computer Science · April 2015

DOI: 10.1007/978-3-319-18111-0_41

3 authors:

Randil Pushpananda
University of Colombo
22 PUBLICATIONS **159** CITATIONS
SEE PROFILE

Ruvan Weerasinghe
University of Colombo
142 PUBLICATIONS **848** CITATIONS
SEE PROFILE

Mahesan Niranjan
University of Southampton
301 PUBLICATIONS **5,900** CITATIONS
SEE PROFILE

Statistical Machine Translation from and into Morphologically Rich and Low Resourced Languages

Randil Pushpananda[1], Ruvan Weerasinghe[1], and Mahesan Niranjan[2]

[1] Language Technology Research Laboratory,
University of Colombo School of Computing, Sri Lanka
{rpn,arw}@ucsc.lk
[2] School of Electronics and Computer Science,
University of Southampton, Highfield, Southampton SO17 1BJ, UK
M.Niranjan@Southampton.ac.uk

Abstract. In this paper, we consider the challenging problem of automatic machine translation between a language pair which is both morphologically rich and low resourced: Sinhala and Tamil. We build a phrase based Statistical Machine Translation (SMT) system and attempt to enhance it by unsupervised morphological analysis. When translating across this pair of languages, morphological changes result in large numbers of out-of-vocabulary (OOV) terms between training and test sets leading to reduced BLEU scores in evaluation. This early work shows that unsupervised morphological analysis using the Morfessor algorithm, extracting morpheme-like units is able to significantly reduce the OOV problem and help in improved translation.

1 Introduction

Recent developments in machine translation are dominated by statistical and machine learning methodologies [1] over rule based approaches [2]. SMT relies on the availability of large corpora of parallel text in the source and target languages. The success of practical machine translation systems such as *Google Translate*[1] and similar systems is somewhat restricted to European languages and Chinese and Arabic in which such large collections of data are available [3]. An additional challenge faced by SMT comes from morphological richness in either the source or target language, or both [4–6]. Morphological modifications amplify the effective vocabulary size at the word and phrase levels resulting in an increased size of corpus needed to estimate their statistics reliably. The challenge is most pronounced when both the source and target languages are morphologically rich, and are minority languages in the sense that there are no readily available large corpora with which SMT systems may be trained.

In this paper, we consider one such language pair, Sinhala and Tamil, the national languages of Sri Lanka. Sinhala is spoken almost exclusively in Sri

[1] https://translate.google.com

A. Gelbukh (Ed.): CICLing 2015, Part I, LNCS 9041, pp. 545–556, 2015.
DOI: 10.1007/978-3-319-18111-0_41

Lanka, while Tamil is found (on a much larger scale) in India as well. Sinhala largely belongs to the Indo-European family of languages while Tamil is from the Dravidian family, mostly found in Southern India. Both are morphologically rich. There are 110 noun word forms and up to 240 verb word forms in Sinhala [7] and about 40 noun forms and up to 240 verb forms in Tamil [8]

Another issue to consider is that, written Tamil and colloquial Tamil differ, and this difference is much more pronounced in the usage of this language in India than in Sri Lanka. This causes particular problems in acquiring parallel corpora to translate between Sinhala and Tamil, a point also noted in [9]. Thus, though some development in natural language processing tools, such as part of speech taggers and morphological analyzers for the Tamil language have been developed in Indian research institutions [10, 11], these are not readily applicable on Sinhala-Tamil parallel corpora that we have collected. Hence we resort to an unsupervised learning approach (see Section 3, Methodology).

There is some early research in SMT between Sinhala and Tamil. One of us [12], showed that a Sinhala-Tamil machine translation is easier than Sinhala-English, and attributed the difference to closer relationships between Sinhala and Tamil due to co-evolution between them that has taken place in Sri Lanka. In that research, 4064 Sinhala and Tamil parallel sentences were used as the training data and 167 Tamil sentences were used for testing. The best BLEU score achieved for that Tamil-Sinhala translation was 13.62% while for English-Sinhala translation it was just 6.18%. In our earlier work [13], we quantified limitations of phrase based statistical translation between Sinhala and Tamil. In particular, we explored the increase in translation accuracy as a function of increasing corpus size. As we expected OOV (unique word) rate is reduced by 8% - 10%, when the parallel dataset size was increased from 5000 to 25000 parallel sentences. However, the error analysis showed that most of the untranslated words were inflections and derivatives. Further, in undergraduate work, attempts were made to translate between Sinhala and Tamil using kernel ridge regression [14], a machine learning method using small corpora of up to 3000 parallel sentences [15, 16]. In this formulation, the mapping between phrases is formulated as a regression with the context in which they appear in a sentence being inputs to the kernel model.

A particular advantage of this language pair is that the primary syntactic structures of both languages are of the Subject-Object-Verb (SOV) form. However their grammars allow substantial differences in word order. We illustrate this in Figure 1. The related English sentences for the given examples of Sinhala and Tamil languages are "We are going to watch a film in a little while", "Altogether there are sixty fruits in that room" and "Asian Development Bank has provided 1250 million rupees" respectively.

Such word / phrase movements have been modeled as regression problems in Ni *et al.* [17] to enhance SMT between grammatically very different languages. We note from the example in Figure 1 that, this may not be a serious issue to consider between Sinhala and Tamil.

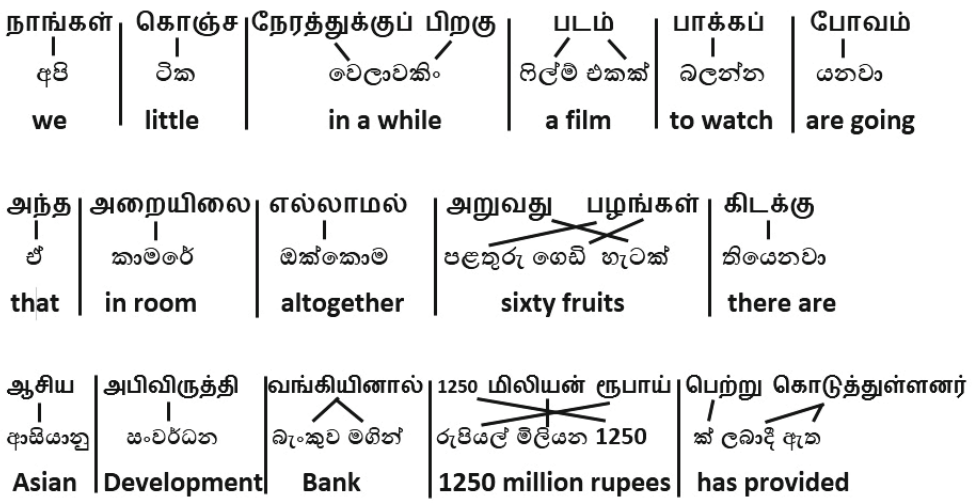

Fig. 1. Example of a pair of Sinhala and Tamil sentences and their relative alignment to the corresponding English sentence taken from the dataset used. Some obvious grammatical errors apparent to readers of Tamil also highlight the challenge of acquiring parallel data for this endeavor.

The use of morphological information integrated into SMT, however, has not been attempted before, both due to a lack of suitable natural language processing tools for Tamil (more specifically, Sri Lankan Tamil) and the datasets used so far being of very limited sizes. The empirical work we report in this paper is a first step in that direction.

Table 1. Examples of some words in common usage in Sinhala with their root (probably) in Tamil

Tamil Word	Sinhala Word	Meaning
அம்மா (/amma:/)	අම්මා (/amma:/)	Mother
அக்கா (/akka:/)	අක්කා (/akka:/)	Elder Sister
பருப்பு (/paruppu/)	පරිප්පු (/parippu/)	Dhal
ஆண்டு (/a:nndu/)	ආණ්ඩු(/a:nndu/)	Government
இடம் (/idam/)	ඉඩම් (/idam/)	Land
வினாடி (/wina:di/)	විනාඩි (/wina:di/)	Minute

In addition there are aspects of Tamil influence on the structure of the Sinhala language. The most significant impact of Tamil on Sinhala has been at the lexical level [18]. Table 1 lists some loan words, of more than a thousand identified as borrowed from Tamil to the Sinhala language [19]. According to [20] people of South India continuously visited Sri Lanka and had close connections with the Sinhala community. Such a close relationship affected the Sinhala language and brought further change to its lexicon.

The remainder of this paper is organized as follows. In Section Two, we give some brief background to statistical translation and morphological analysis that is relevant to this work. In Section Three, we discuss methodological details. Section Four describes experimental evaluations carried out, and we conclude with a discussion of results obtained in Section Five.

2 Background

Translation between two morphologically rich languages is still uncommon. However translating from English to a morphologically rich language and vice versa are widely studied problems. According to the literature, various approaches have been applied for the translation between morphologically rich languages. Most of the researchers have used morphological analyzers and part of speech (POS) taggers to integrate the morphological information to machine translation research.

Popović *et al.* [21] have showed that there are some significant improvements to be achieved by considering morpho-syntactic information even considering only the base forms of the Serbian language when translating from Serbian to English. Also translating from English to Serbian can be improved by removing some of the articles in English. However, it clearly shows that translation in the English to Serbian direction has higher word error rate rather than in the other direction. This proves the difficulty of translating into a morphologically rich language due to the large number of inflections. Similarly, Oflazer *et al.* [22] have investigated phrase based SMT from English to Turkish. They have used lexical morphemes instead of surface morphemes. Results obtained by them show that somewhere in between full word forms and fully morphologically segmented representations provide a significant BLEU score improvement. Nießen and Ney [23] have shown the importance of morphology for scarcely resource languages. Popvic and Ney [24] have proposed two methods to improve the quality of translation from Serbian, Spanish and Catalan languages to English by using POS tags, words stems and suffixes. Their work resulted in a significant reduction of error rates for Serbian, Spanish and Catalan languages. Segalovich [25] showed that an algorithm which can be used to get the morphology of wide lexical coverage using only a limited dictionary.

The above studies investigated translating from English to a morphologically rich language or from a morphologically rich language to English. There are only a limited number of research carried out for translation between two morphologically rich languages without having full morphological analyzers and POS taggers. In such a case an unsupervised morphology learning preprocess to SMT is one of the approaches to be explored. Virpioja *et al.* [26] have applied the Morfessor [27] algorithm to extract the morpheme-like units in an unsupervised manner. They have shown that longer n-grams and longer phrase lengths result in better values for morph-based translations. Even though BLEU score values were slightly lower than the word based approach, it showed promising results for the two morphologically rich languages.

However, there has been no integration of morphology into SMT reported in the literature for the Sinhala-Tamil language pair. Therefore this empirical research is expected to be helpful to identify better morphological approaches to build a successful MT system for translating between two morphologically rich and resource poor languages.

3 Methodology

In this research we have used Morfessor, an unsupervised learning algorithm, to find morpheme-like units of the source and target languages in order to train the language and translation models. Since Morfessor Categories-MAP algorithm [28] has a better segmentation accuracy and handles OOV words in the training data [26], we have used it in our work. Using this approach words have been divided as multiple prefixes followed by stem(s) and multiple suffixes. In rare cases we have found some multiple stems as well.

First we trained the Sinhala and Tamil datasets separately using Morfessor and extracted morpheme-like units as shown in the Figure 2.

සහෝදරියන්ට (To sisters)	සහ/PRE + ෝ/STM + දර/SUF + ි/SUF + ය/SUF + න්/SUF + ට/SUF
කෙටිකාලීනව (short period)	කෙට්/PRE + කා/PRE + ලී/STM + න/SUF + ව/SUF
அகதிகளின் (Refugees)	அ/PRE + க/STM + தி/SUF + களின்/SUF
அகமதாபாத்தில் (In Ahmedabad)	அ/PRE + கம/STM + தா/SUF + பா/SUF + த்/SUF + தில்/SUF

Fig. 2. Examples of unsupervised morphological decomposition

Then we performed three sets of experiments, one with a word based (Baseline system) and two others with two different morphological representations (fully morpheme-like and semi morpheme-like segmentation systems) for the Sinhala-Tamil language pair.

Baseline System. In this experiment, we have used the standard phrase-based SMT modelling approach where words were used as the smallest unit. We have done this experiment to compare performance against the two morph-based approach. A sample parallel sentence from the data is shown below. Here Tamil (TA) sentence length is 7 and Sinhala (SI) sentence length is 10

TA: அவர் கண்ணீர் மல்கிய கண்களுடன் தனது மனைவியை பார்த்துக்கொண்டிருந்தார்
SI: ඔහු කඳුළු පිරුණු දෙනෙතින් යුතුව තම බිරිඳ දෙස බලා සිටියේය

To develop the baseline system, the open source SMT system MOSES [29] was used with GIZA++ [30] using the standard alignment heuristic grow-diag-final

for word alignments. Language models were trained using the Stanford Research Institute language Modeling (SRILM) toolkit [31] with Kneser-Ney smoothing. The systems were tuned using a small extracted parallel dataset (500 sentences) with Minimum Error Rate Training (MERT)[32] and then tested with a randomly extracted test dataset (10% of training data). Finally, the Bilingual Evaluation Understudy (BLEU) [33] evaluation metric was used to evaluate the output produced by the translation system.

Fully Morpheme-Like Segmentation System. In this method morpheme-like units used as the smallest unit and phrase based SMT modelling approach was used similar to the baseline system. As in Figure 2, resulting surface morphemes consist of tags such as Prefixes (PRE), Stems (STM) and Suffixes (SUF). However before training the translation model and the language model, we have removed these tags from the data. Then words in the parallel sentences (training, tuning, testing) and monolingual corpus were replaced with these morpheme-like units. A sample of the split morpheme-like parallel sentence is shown below. Here the Tamil (TA) sentence length is 19 and Sinhala (SI) sentence length is 21

TA: அவர் | கண்ண ீ ர் | ம ல் கி ய | கண் களுடன் | தன து | மனைவி யை | பார்த்த ா க் கொண்டிருந்த ார்
SI: ඔහු | කඳු ළු | පි රු ණු | දෙ නෙ ත ින් | යුතු ව | තම | බි රි ඳ | දෙස | බලා | සිටි යේ ය

Then as mentioned in the baseline system, training, testing and tuning were done. Finally the evaluation was done after performing some post processing. In the post processing stage, the longest matching morpheme-like units were merged to extract readable translated sentences.

Semi Morpheme-Like Segmentation System. In this approach we have combined all the prefixes and stems together and separately merged the suffixes. Similarly, we have built the translation model and language model as before. A sample parallel sentence of this form is shown below. Here the Tamil (TA) sentence length is 12 and Sinhala (SI) sentence length is 15

TA: அவர் | கண்ண ீர் | மல் கிய | கண் களுடன் | தன து | மனைவியை | பார்த்துக்கொண்டி-
ருந்த ார்
SI: ඔහු | කඳු ළු | පිරුණු | දෙනෙ තින් | යුතු ව | තම | බිරි ඳ | දෙස | බලා | සිටියේ ය

Similar to the fully morpheme-like segmentation approach, evaluation was done after post-processing the resulting output. Finally all the results were compared with the baseline system.

4 Experimental Conditions

4.1 Data

We have conducted our experiments for the Sinhala-Tamil language pair. Since Sinhala and Tamil parallel data is limited, we have used two methods to collect the parallel data. The first approach was identifying Sinhala-Tamil parallel documents such as magazines, books, articles, etc. Then we checked the availability of parallel data in electronic format. Most of the documents were available in electronic form but in pdf format not encoded in Unicode. Therefore we had to convert proprietary encoding into Unicode. However we had to align the sentences manually since they were not pre-aligned. Most of the sentences were aligned in a one to many form and some were not aligned at all. Also we have found a large number of figures and tables inside these documents which made the sentence alignment harder.

The second approach to collect parallel data was by translating sentences available electronically in one language to the other with the help of professional translators. We first extracted Sinhala sentences from the *UCSC* [2] *10M words Sinhala Corpus* [34] with word lengths between 8 and 12. Professional translators then translated these Sinhala sentences to Tamil. From both these approaches, we managed to collect 25,500 Sinhala-Tamil parallel sentences. Detailed statistics of the parallel corpus collected are given in Table 2.

Table 2. Characteristics of parallel dataset

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

We used the *UCSC 10M words Sinhala Corpus* to build the Sinhala language model. The characteristics of the Sinhala corpus is shown in Table 3.

Table 3. Characteristics of Sinhala Monolingual Corpus

Language	Characteristics		
	Total Words	Unique Words	Sentences
Sinhala	10,142,501	448,651	850,000

4.2 Experiments and Results

As mentioned in the section 3, we have carried out three sets of experiments separately. All the experiments were done in the Tamil to Sinhala translation direction. Fully morpheme-like segmentation and semi morpheme-like segmentation were done repeatedly for three different language models (3-gram, 5-gram

[2] University of Colombo School of Computing.

and 7-gram) without changing the default phrase length. The word-based baseline approach was carried out only for the default settings (i.e. phrase length: 7 and 3-gram language model). Results obtained for the experiments are shown in Table 4.

Table 4. BLEU Score values of the word-based, fully segmented and semi segmented approaches

Description	Word Based	Fully Segmented			Semi Segmented		
	3-gram	3-gram	5-gram	7-gram	3-gram	5-gram	7-gram
BLEU Score (%)	12.99	8.50	12.06	12.53	9.7	10.68	10.29

By comparing the columns in Table 4, we can clearly see that the word-based baseline system gives better BLEU score results overall. However, when we consider the fully-segmented approach, we can see that the BLEU score values increases while increasing the language model size upto 7-gram. According to Table 4, the BLEU improvement rate of the fully-segmented model has been reduced when increasing the language model size from 3 to 7. When considering the semi-segmented approach, results of the 3-gram language model gave better results than the fully-segmented approach. However when the language model size increases upto 7-gram, 7-gram semi-segmented approach resulted in a lower BLEU score value compared to the 5-gram semi-segmented approach.

Since the semi-segmented approach resulted in lower values compared to the fully-segmented approach, further investigations were conducted only using fully-segmented approach. Further investigations were done by changing the maximum phrase length size to 10 in both 5-gram and 7-gram language models. The evaluation results are shown in Table 5.

Table 5. BLEU score (%) values obtained for the experiments done with 5-gram and 7-gram language model and maximum phrase length 10

Description	Fully Segmented (Phrase Length:10)	
	5-gram LM	7-gram LM
BLEU Score(%)	12.15	13.11

Finally the best BLEU score resulted from the fully-segmented approach with language model size 7-gram and maximum phrase length size 10. However, it is not a significant improvement compared to the results of the baseline system. When we consider the translated output, we can rarely see any untranslated words, unlike in the baseline system. Even though the untranslated words are rare, we could achieve only lower BLEU score values for the fully-segmented approach. Visual inspection of the translated output clearly sehows that the first half of sentences have been translated more accurately rather than the second half of sentences. The Table 6 shows the evaluation of the first and second half

of the sentences in 3-gram and 7-gram language models with maximum phrase length 7.

Table 6. Evaluation of first and second half of the sentences resulted in 3-gram and 7-gram language models

Description	BLEU Score(%)	
	3-gram Language Model	7-gram Language Model
Overall	8.50%	12.53%
First Half	9.93%	17.28%
Second Half	3.21%	4.31%

According to Table 6, we can clearly see that the average BLEU score value of the first half of the sentences are much higher than those of the second half of the sentences. Table 7 shows an average word/morpheme-like unit length of the sentences including the maximum and minimum sentence lengths. According to Table 7, we can clearly see that the Tamil morpheme-like sentence length is 3 times larger than the word based sentence length whereas for Sinhala, it is only twice as large.

Table 7. Average and maximum sentence lengths of each word based and morpheme based sentences. Word based approach words used as the smallest unit and morpheme based approach morpheme-like units (MOR) considered as the smallest unit

Description	Sentence Length	
	Words	MOR
Average (Tamil)	10	27
Maximum(Tamil)	23	60
Average (Sinhala)	11	22
Maximum (Sinhala)	27	59

5 Discussion and Conclusions

Experimental results support the suggestion that integrating morphological information into SMT is a way around the data sparseness issue for language pairs that are morphologically rich. However, in the comparisons we make, the BLEU scores are not significantly higher than those of the baseline. In our earlier study [13]), we noted that the traditional word-based approach was unable to translate 25% of words in the test set, and out of this 68% was owing to words in the test set being OOV. This suggests that the rest of the words (32%) are untranslated even when they were present in the training set. Morphological analysis via unsupervised learning was able to reduce this to less than 1% of the total input words, which is a significant result observed in the experiments.

As seen in Table 7, as larger words in the text gets decomposed into smaller morpheme-like units, sentence lengths increase. Since phrase-based systems work better with short word alignments, this length bias introduced by the morph decomposition needs to be improved on. In the experiments we conducted, the longest matching morphemes were merged as words. We will explore ways of correcting the resulting errors by post-processing methods.

Another important observation we made is the influence of errors in the training dataset. In morphologically rich languages, a certain amount of variability in suffixes and merging of words into compounds is tolerated. Writers are often not consistent in the way they use such variations and do not stay within strict grammatical rules of the language. In our database we see this difficulty in sentences that have been translated from a Sinhala original to Tamil by translators. Manually cleaning the training data is important to address this issue.

In future work, we will also concentrate on enhancements to the morph decomposition approach by focusing on suffixes, as in both languages the main morphological modifications are in this part. We believe the decomposition algorithm could be improved to allow supervised segmentation to achieve this.

Acknowledgments. The authors would like to acknowledge the National Research Council (NRC), ICT Agency and LK Domain Registry of Sri Lanka for funding this research. They are also indebted to the research team members of the Language Technology Research Laboratory of the University of Colombo School of Computing, Sri Lanka, for assisting in numerous ways.

References

1. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2009)
2. Chéragui, M.A.: Theoretical Overview of Machine Translation. In: Proceedings ICWIT, p. 160 (2012)
3. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: MT Summit, vol. 5 (2005)
4. Koehn, P., Hoang, H.: Factored Translation Models. In: EMNLP-CoNLL, pp. 868–876 (2007)
5. Goldwater, S., McClosky, D.: Improving Statistical MT through Morphological Analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 676–683. Association for Computational Linguistics (2005)
6. Davis, E.H., Lavie, P.A., Vogel, S.: Integration of Morphology into Statistical Machine Translation (2008)
7. Welgama, V., Herath, D.L., Liyanage, C., Udalamatta, N., Weerasinghe, R., Jayawardana, T.: Towards a Sinhala Wordnet. In: Proceedings of the Conference on Human Language Technology for Development (2011)
8. Lushanthan, S., Weerasinghe, R., Herath, D.: Morphological Analyzer and Generator for Tamil Language. In: Proceedings of the 14th International Conference on Advances in ICT for Emerging Regions, Colombo, Sri Lanka, pp. 190–196 (2014)

9. Germann, U.: Building a Statistical Machine Translation System from Scratch: How much bang for the buck can we expect? In: Proceedings of the Workshop on Data-Driven Methods in Machine Translation, vol. 14, pp. 1–8. Association for Computational Linguistics (2001)
10. Parameshwari, K.: An Implementation of Apertium Morphological Analyzer and Generator for Tamil. An E-Journal of Language in India (2011), http://www.languageinindia.com
11. Anand Kumar, M., Dhanalakshmi, V., Soman, K., Rajendran, S.: A Sequence Labeling Approach to Morphological Analyzer for Tamil Language. IJCSE) International Journal on Computer Science and Engineering 2, 1944–195 (2010)
12. Weerasinghe, R.: A Statistical Machine Translation Approach to Sinhala-Tamil Language Translation. Towards an ICT Enabled Society 136 (2003)
13. Pushpananda, R., Weerasinghe, R., Niranjan, M.: Sinhala-Tamil Machine Translation: Towards better Translation Quality. In: Proceedings of the Australasian Language Technology Association Workshop 2014, Brisbane, Australia, pp. 129–133 (2014)
14. Wang, Z., Shawe-Taylor, J., Szedmak, S.: Kernel Regression Based Machine Translation. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, pp. 185–188. Association for Computational Linguistics (2007)
15. Jeyakaran, M.: A Novel Kernel Regression Based Machine Translation System for Sinhala-Tamil Translation. Unpublished BSc Thesis (2011)
16. Sakthithasan, S.: Statistical Machine Translation for Sinhala and Tamil. Unpublished BSc Thesis (2010)
17. Ni, Y., Saunders, C., Szedmak, S., Niranjan, M.: Exploitation of Machine Learning Techniques in Modelling Phrase Movements for Machine Translation. Journal of Machine Learning Research 12, 1–30 (2011)
18. Karunatilaka, W.: Link. Godage International Publishers, Sri Lanka (2011)
19. Coperahewa, S., Arunachalam, S.: A Dictionary of Tamil Word in Sinhala, vol. 2. Godage International Publishers, Sri Lanka (2011)
20. Chandralal, D.: Sinhala, vol. 15. John Benjamins Publishing (2010)
21. Popović, M., Vilar, D., Ney, H., Jovičić, S., Šarić, Z.: Augmenting a Small Parallel Text with Morpho-Syntactic Language Resources for Serbian-English Statistical Machine Translation. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts, pp. 41–48. Association for Computational Linguistics (2005)
22. Oflazer, K., El-Kahlout, I.D.: Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 25–32. Association for Computational Linguistics (2007)
23. Nießen, S., Ney, H.: Statistical Machine Translation with Scarce Resources using Morpho-Syntactic Information. Computational Linguistics 30, 181–204 (2004)
24. Popovic, M., Ney, H.: Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In: LREC (2004)
25. Segalovich, I.: A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In: MLMTA, CiteSeer, pp. 273–280 (2003)
26. Virpioja, S., Väyrynen, J.J., Creutz, M., Sadeniemi, M.: Morphology-Aware Statistical Machine Translation based on Morphs Induced in an Unsupervised Manner. In: Machine Translation Summit XI, pp. 491–498 (2007)

27. Creutz, M., Lagus, K.: Unsupervised Models for Morpheme Segmentation and Morphology Learning. ACM Transactions on Speech and Language Processing (TSLP) 4, 3 (2007)
28. Creutz, M., Lagus, K.: Inducing the Morphological Lexicon of a Natural Language from Unannotated Text (2005)
29. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: MOSES: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (2007)
30. Och, F.J., Ney, H.: The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics 30, 417–449 (2004)
31. Stolcke, A., et al.: SRILM-An Extensible Language Modeling Toolkit. In: INTERSPEECH (2002)
32. Och, F.J.: Minimum Error Rate Training in Statistical Machine Translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1, pp. 160–167. Association for Computational Linguistics (2003)
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
34. Weerasinghe, R., Herath, D., Welgama, V., Medagoda, N., Wasala, A., Jayalatharachchi, E.: UCSC Sinhala Corpus - PAN Localization Project-Phase I (2007)