

# A Study on the Utility of Hierarchical Phrase-Based Model for Low Resource Languages

S.Yashothara<sup>1</sup> and R.T.Uthayasanker<sup>2</sup>

Department of Computer Science and Engineering,  
University of Moratuwa,  
Sri Lanka.

<sup>1</sup>yashoshan@gmail.com, <sup>2</sup>rtuthaya@cse.mrt.ac.lk

**Abstract.** This paper addresses the Hierarchical Phrase Based (HPB) models which are used in development of different Statistical Machine Translation (SMT) Systems for four low resourced South Indian languages. Currently, South Asian languages are dominantly translated using traditional statistical and neural machine translation approaches. South Asian languages lack necessary natural language resources and tools hence classified as low resourced languages. Any SMT System needs large parallel corpora for exact performance. So, non-availability of corpora limits the success achievable in machine translation to and from those languages. Compared to English, South Asian languages are morphology rich and commonly use different sentence structure. The structure of a sentence is Subject-Verb-Object in English while Subject-Object-Verb in most of the South Asian languages. As South Asian languages are low resourced, it is difficult to get a good order of sentences when traditional Statistical Machine translation is used. They can only be reordered using distortion reordering model, which is independent of their context. To overcome this problem hierarchical phrase model translation which uses grammar rules formed by the Synchronous Context Free Grammar can be used. This paper considers English to Tamil, Tamil to English, Malayalam to English, English to Malayalam, Tamil to Sinhala and Sinhala to Tamil translations. At the end, automatic evaluation of system is performed by using BLEU as evaluation metrics. Hierarchical phrase based model shows better result compared to traditional approach between Tamil-English and Malayalam-English pairs. For Sinhala to Tamil, it achieves 11.18 and 10.73 for vice-versa. Moreover, the system could be improved by adding some rules.

**Keywords:** Hierarchical Phrase Based model, Statistical Machine Translation, Parallel corpus, Natural Language Processing.

## 1 Introduction

With the rebellion of the internet, people have more opportunities to go global. However, communication is made more challenging due to differences in language. Even though, English is widely accepted as an official language in many multilingual countries like Sri Lanka and India, it cannot be assured that everyone knows it. Therefore, translation plays a major role. Currently, South Asian languages are dominantly translated using traditional statistical and neural machine translation approaches. But, most of the South Asian languages are low resourced due to lack of necessary natural language resources and tools.

Compared to English, South Asian languages are morphology rich and common-

ly use different sentence structure. The structure of a sentence is Subject-Verb-Object in English while Subject-Object-Verb in most of the South Asian languages. In most of the South Asian languages, there is no difference between capital and lowercase letters. The difference between colloquial and formal form of these languages is also much greater compared to English. The above discussion emphasizes the differences between the source languages and the target languages. This research focused on Tamil to English, English to Tamil, Malayalam to English, English to Malayalam, Tamil to Sinhala and Tamil to Sinhala.

As South Asian languages are low resourced, it is difficult to get a good order of sentences (because of sub-phrases) when traditional Statistical Machine translation is used. They can only be reordered using distortion reordering model, which is independent of their context. Also learning phrases longer than three words barely improve the translation because such phrases are infrequent in the corpora due to data sparsity. Data sparsity is very likely that we will not see all of the words at training time. But we have a lot of unordered text which are similar in some sense.

To overcome the above issues, we adopted hierarchical phrased based machine translation [1][2] which belongs to one of the current leading and promising statistical machine translation approaches in this research. Phrase-based translation is expanded by hierarchical phrase based translation by allowing phrases with gaps and modeled as Synchronous Context-Free Grammar (SCFG) [3]. Hierarchical model brings sub-phrases into existence in order to remove the problems associated with phrase-based MT. Although global reordering SCFG is captured by model, the reordering does not explicitly introduce the model to restrict word order. As hierarchical phrase translation uses the nonterminal symbols, Lexicalized reordering models used in traditional machine translation cannot be applied directly to hierarchical phrase based translation [3].

An important concern with hierarchical based translation is the size of the model on which training is carried out, which is usually several times larger than the trained phrase based counterpart from the same dataset. But, this heads to over generation search errors and a slow decoder [4]. In this work, the key focus is on hierarchical model's usage in Statistical Machine Translation (SMT) for the South Asian languages, focused on the expression of Chiang et al's (2005) work.

HPB SMT [3] combines the strength of a rule-based and a phrase-based machine translation system. It constructs trees by automatically extracting a SCFG from the training corpus. The basic unit of HPB SMT is hierarchical rules which are extracted by Context Free Grammar according to the phrase based model [5]. So, hierarchical rules have the strength of learning sentence reordering without a separate reordering model.

We conducted experiments with hierarchical phrase based translation using Moses, for the translations between Tamil-English, Malayalam-English and Tamil-Sinhala languages, and compared the results with traditional phrase-based models with the same corpora. We have selected Tamil-Sinhala pair of languages to check the hierarchical model, which has the same sentence structure.

## 2 Literature Review

This section reviews the literature about adding hierarchical phrase based model Statistical Machine Translation system and existing Machine Translation systems for Tamil, English, Malayalam and Sinhala languages.

As the first step, hierarchical phrase based model for SMT for different European languages is proposed by Chiang. Their experiments were on Mandarin-to-English translation and claimed HPB SMT system achieves an absolute improvement of 0.02 over the traditional SMT (7.5% relative), without using any additional training data [3]. Mahsa Mohaghegh and Abdolhossein Sarrafzadeh adopted this method for the translation between English and Persian languages. They have compared Moses tool kit and Joshua tool kit by dividing corpus into sections of 20K, 30K, 40K, and 50K sentences. The best result claimed in the paper is 4.5269 NIST and 0.3708 BLEU using the Joshua based system trained on 50K corpus [6].

One of the early works on hierarchical phrase based model in SMT for south Asian languages was by Jawaaid et.al who examined between English and Urdu. They experimented using the Moses SMT system and presented an Urdu aware approach based on reordering phrases in syntactic parse tree of the source English sentence [7]. Nadheem Khan et al. have focused on English to Urdu HPM SMT. 6596 sentences parallel corpus is used for training. The k-fold cross validation method was used for sampling of the corpus. Here k=5 was selected by taking 4/5 of the total corpus as training and 1/5 as tuning and test set for experiment on all folds. Highest percentage of result is 29% in the experiment [8].

Various works with different approaches have been proposed for translation between Tamil-English, Malayalam-English and Tamil-Sinhala languages. The following fragment will provide the significance of machine translation and will identify a place where a new contribution could be made for those languages by analyzing published information in the area of machine translation.

Ulrich Germann [9] conveyed his experience with building a SMT system for translation between Tamil and English from scratch, including the creation of a small parallel Tamil-English corpus. Following this research, there are several other researches [10], [11] using traditional SMT. Loganathan developed SMT system by integrating morphological information. He separated the morphological suffixes to improve the quality of traditional phrase based model [12]. Anandkumar et al. adopted factored SMT system to handle the morphologically fluent Tamil sentences. They applied the manually created reordering rules to the syntactic trees for rearranging the phrases in English. This improves the performance in local distance sentences and already available sentences in the training corpora [13]. But long distance reordering and new sentence reordering are not handled in these approaches.

First effort ‘Rule Based translation system’ reported in the translation from Malayalam to English [11]. But, development of rule-based systems requires more cost, time extensive linguistic rules and it sometimes fails to find good translation due to search errors during the decoding process. Sebastian et al. proposed a SMT approach by adding some pre-processing and post processing steps. Alignment model is

increased by adding the parts of speech information into the bilingual corpus and removing the inappropriate alignments from the sentence pairs. Corpus is pre-processed by suffix and stop word elimination techniques. They have used order conversion rules to resolve the structural difference between English and Malayalam languages [15]. But, adding rules to translation also faces problems such as high cost in formulating rules and conflicts when the numbers of rules increase.

Considering local languages of Sri Lanka (Sinhala -Tamil) very minimal of researches have been carried out to date. As a first attempt, Ruvan Weerasinghe proposed a basic SMT approach to Sinhala and Tamil languages. After testing with multiple translation models, they have achieved a best BLEU score of 0.1362 for this task [16]. Following that work, S. Sripirakas et al. proposed translation system which has been implemented with the preparation of parallel corpora from parliament order papers. They demonstrate only the preliminary system which runs both directions of Tamil and Sinhala languages [17]. There are some similar approaches [18], [19], [20], [21] by using SMT. But there have been no efforts to translate between Tamil and Sinhala languages using hierarchical phrase based statistical machine translation. But, traditional statistical based machine translation (SMT) mostly fails to produce quality output for long sentences.

By having a look at the work above, it is clear that there is not a single proposed work in hierarchical phrase based statistical machine translation for Tamil, Malayalam and Sinhala.

### 3 Evaluation

This section discusses the training, tuning and testing of different model components. The evaluation was carried out on Ubuntu 16.00 running on Intel Core i5 machine with 2GB of RAM and 500GB of Hard disk space between Tamil-English, Malayalam-English and Tamil-Sinhala.

#### 3.1 Dataset

We used the IIIT-Hyderabad (International Institute of information Technology) parallel corpus for Tamil-English and Malayalam-English languages. They have corpora of eleven languages. Size of each corpus is about 3 million words. Texts in each corpus are categorized under aesthetics, mass media, social science, natural science, commerce and translated materials. The corpora were prepared by several organizations under the funding from MoIT (Ministry of Information Technology formerly Department of Electronics), Government of India. Its bilingual resources consist of roughly about 50,000 sentences for all the available languages [22]. The corpora are already sentence aligned. Here we clean this corpus for making it completely compatible.

The main source of the parallel corpus of Sinhala-Tamil languages is government official documents. The documents collected from government institutions

were hard copies and some were of a single source. They are generally translated manually with the aid of human translators. We digitalized those written documents into text files by crowdsourcing. The typed documents were sentence aligned with the manual inference. Its bilingual resources consist of about 22,000 sentences for all the available languages. Further details about parallel data are given in Table 1.

**Table 1:** Complete Statistics of Parallel Corpus (In Sentence)

	<b>Tamil- English</b>	<b>Malayalam- English</b>	<b>Tamil- Sinhala</b>
Training	48,000	48,000	20,000
Development	1,500	1,500	1,500
Testing	500	500	500

The target language corpus in above parallel corpus is used in the development of language model for this study work.

### 3.2 Experimental setup

The experiments were conducted to check the applicability of hierarchical phrase based model in translation between morphologically rich languages and morphologically rich and poor languages. English-Tamil, Malayalam-English pair of translations were selected for the experiment of translation between morphologically rich and poor language, Tamil and Sinhala languages are chosen for the experiment of translation between morphologically rich languages.

As the initial step of the experiments, the obtained data was tokenized using customized scripts and standard Moses [23] filtration was utilized to confirm that the sentences with extreme length ratio difference were removed effectively. English language corpus was followed by lowercasing by the script being supplied with the Moses decoder [23]. This training data was used for word alignment. Moses [23] was run using Koehn’s training scripts. In our work additional switches like hierarchical and glue grammar were also used in training command as the experiments were carried out with the HPB model.

For the other parameters, the default values were used i.e. 3-gram language model and maximum phrase length= 6. Giza++ [25] was used for the word alignment with ‘grow-diag-final-and’ as the summarization heuristics. Lmplz [24] was used for the language modeling. 3-gram Language Models (LMs) were created. The featuring weights were tuned using Minimum Error Rate Training (MERT) on 100 best translations [23]. A set of 1500 randomly selected sentences were used for tuning. Decoding was done using the state-of-the-art Moses using cube pruning techniques with stack size of 5000 and the maximum phrase length of 5 [26]. The testing phase was completed by using the Moses decoder. The testing was carried out in the same way for all the language pairs. For the comparison the results of HPM SMT, we have done tradi-

tional SMT approach also to the same data set. Traditional SMT approach for the same data set was also used for the comparison of the HPM SMT results.

The output of the system was evaluated using Bilingual Evaluation Understudy (BLEU) [27]. The system was evaluated on 500 randomly selected sentences / phrases, where the letter headers and footers were added as comma separated phrases for testing, to ensure that the score of a single sentence no longer depends on a single or very little amount of words.

### 3.3 Results

The evaluation scores of the aforementioned three language pairs in both the directions and the sample translations from the developed HPM SMT are described in this section. In each language pair we trained the SMT with and without HPM and evaluated its translation quality by measuring the BLEU score of the translation of test data set. Even though, these language resources are sparse, we have achieved much better BLEU score for the entire set of language pairs. The scores of six different experimental setups are tabulated in Table 2. A comparison of the developed hierarchical phrase based translation system with the traditional phrase based system was also carried out for the same dataset.

It can be noted from Table 2, that the hierarchical phrase based model system got better BLEU scores compared to the traditional Phrase based model approach for Tamil to English, English to Tamil, Malayalam to English and English to Malayalam. While those differences are less since the dataset size is small, the percentage of the difference is high. These results show that usefulness of hierarchical phrase based model is significant when there is different in sentence structure between the languages getting translated. Nevertheless, for the translation of Tamil to Sinhala and Sinhala to Tamil, it could be noticed from Table 2 that the traditional phrase based model system got better BLEU scores compared to the hierarchical phrase based model approach. The main reason behind this is that both Tamil and Sinhala language share same sentence structure and morphologically rich. Further the Tamil-Sinhala corpus is the smallest among the three which causes sparseness in training data. HPM is sensitive to sparse data and that could have further reduced the translation quality in this case. These observations show that the HPM is most useful in language pairs varied by sentence structure but would affect the quality of the translation if the languages share the same sentence structure.

**Table 2:** Comparison of BLEU evaluation score with traditional Phrase based model

	BLEU Score		Differentiation
	Traditional SMT	Hierarchical Model	(%)
Tamil to English	3.16	3.42	8.23
English to Tamil	1.17	1.73	47.863

Malayalam to English	4.22	4.40	4.26
English to Malayalam	2.88	3.310	14.93
Tamil to Sinhala	*14.88	11.18	(-20.16)
Sinhala to Tamil	*13.61	10.73	(-21.17)

The results also show that BLEU score increase is higher from English to Tamil or Malayalam compared to the other direction. As we know Tamil & Malayalam are morphologically richer than English. In these cases HPM leverages the morphological divergence between these languages in its favour. Also from the results, it can be noted that the translation from morphologically rich languages (Tamil, Malayalam) to morphologically poor languages (English) gives better BLEU score in traditional SMT and HPM SMT compared to other way around. Even though Sinhala is a morphologically rich language, the translation from Tamil to Sinhala shows higher results as Tamil language is morphologically richer than the Sinhala language. These observations show that the translations from morphologically rich languages to morphologically poor languages gives better result compare to other direction.

Also, English to Tamil got the highest percentage of increase in BLEU score due to HPM compared to traditional SMT (47%), and Sinhala to Tamil got the highest decrease in BLEU score percentage (21%). From the results, it can be observed that, the translations from morphologically poor languages (English) to morphologically rich languages (Tamil, Malayalam) give more improvement using HPM model. So, usefulness of HPM is significant when divergence of morphology and divergence of sentence structure.

Figure 1 shows how the decoder performs translations of the test dataset using the chart decoder for hierarchical phrase based model. For the input Tamil sentence “ஆரம்பத்திலே சிறிய உடற்பயிற்சி செய்யுங்கள்.”, the sentence is translated as “Start with light exercise”.

```

Translating: <s> அழகத்திலே சிறிய உறையிருக்கிறார்கள் . </s> ||| [0,0]=X (1) [0,1]
=X (1) [0,2]=X (1) [0,3]=X (1) [0,4]=X (1) [0,5]=X (1) [0,6]=X (1) [1,1]=X (1)
[1,2]=X (1) [1,3]=X (1) [1,4]=X (1) [1,5]=X (1) [1,6]=X (1) [2,2]=X (1) [2,3]=X
(1) [2,4]=X (1) [2,5]=X (1) [2,6]=X (1) [3,3]=X (1) [3,4]=X (1) [3,5]=X (1) [3,6]
=X (1) [4,4]=X (1) [4,5]=X (1) [4,6]=X (1) [5,5]=X (1) [5,6]=X (1) [6,6]=X (1)

0 1 2 3 4 5 6
1 1 15 11 5 18 0
1 1 121 52 70 0
1 14 200 200 0
12 52 200 0
37 199 0
59 0
1
BEST TRANSLATION: 6701 S -> S </s> :0-0 : term=1-1 : nonterm=0-0 : c=-1.310 cor
e=(0.000,-1.000,1.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000) [0..6] 5575 [total=-
0.943] core=(0.000,-7.000,8.000,-3.526,-12.089,-0.226,-11.022,3.000,-19.052)
Start with light exercise .
Line 0: Additional reporting took 0.004 seconds total
Line 0: Translation took 0.046 seconds total
Translation took 0.022 seconds
Name:moses VmPeak:505216 kB VmRSS:332940 kB RSSMax:350016 kB u
ser:5.944 sys:0.432 CPU:6.376 real:6.532

```

**Fig. 1.** Working of Hierarchical Phrase based decoder for Tamil to English translation.

Some examples of translation generated by translation system developed in this study are provided in Table 3. Two examples of each different translation have been listed here. Some of the examples are not perfect translations. This may occur since the South Asian languages are rich in morphology compared to English, there may be noise in training data, out of vocabulary, misordering of words, wrong alignment of phrases, inappropriate translation to the context and harder sparse-data problems due to vocabulary that combines words from various sources. However, there are some examples below which show hierarchical phrase based model helps to reorder the sentences.

**Table 3:** Some examples of translation generated by the translation system developed in this study.

	Input	Output
<b>Tamil to English</b>	சைட்டிகா மற்றும் சிலிப்புடிஸ்க் நோயாளி இந்த பயிற்சி செய்யாதீர்கள்	The patients of sciatica and slip disk should avoid its practice
	பூங்காவிற்கு பைக் எடுத்துச்செல்ல அனுமதியில்லை.	The a rate of a take her அனுமதியில்லை
<b>English to Tamil</b>	Drink plenty of water	நன்றாக தண்ணீர் குடியுங்கள்
	Chew the sugar-free chewing gum	சர்க்கரை இல்லாத சூயிங்கம் மெல்லவேண்டும்
<b>Malayalam to English</b>	പതിവായിട്ട് ദന്തനിരീക്ഷണം ചെയ്യണം .	Get the teeth checked-up regularly
	നെഓറാ താഴ്വര ദേശീയ ഉദ്യാനത്തിൽ അനേകം	Neora Valley National Park is Approximately 150 species of

	ഓർക്കിഡുകൾ ഉണ്ട് .	orchids are found
<b>English to Malayalam</b>	New Digha is the new tourist spot of Digha	ചിൻഡിയിലെ കമരൂനാഗ് . ഗിവഗുഹ തുങ്ങിയവ കാണേണ്ടതാണ് .
	There is ' Forest Hut ' in Sanarali 3 kms ahead	3 കി.മീ. സനാരിയിൽ ഫാം റെസ്റ്റ് ഹട്ട് ഉണ്ട് .
<b>Tamil to Sinhala</b>	கலந்துரையாடல் நிகழ்வு செயலாளரின் தலைமையில் இடம்பெற்றது	සංවාද සිද්ධිය ලේකම්ගේ ප්‍රධානත්වයෙන් පැවැත්විණ .
	பணிக்கான கணக்குப் பிரிவு	කාර්යය ගිණුම් අංශය
<b>Sinhala to Tamil</b>	ජල පොම්ප අළුත්වැඩියා අංශය	நீர் திருத்துதல் பிரிவு
	ජල සැපයුම් අංශය	நீர் வழங்கல் பிரிவு

#### 4 Conclusion

In this work we have explored one of the important but relatively less addressed research problems. This is the first work on developing a hierarchical phrase based SMT for English, Tamil, Malayalam and Sinhala languages. In this work hierarchical phrase based model was applied in the translations of Tamil to English, English to Tamil, Malayalam to English, English to Malayalam, Tamil to Sinhala and Sinhala to Tamil. The comparison between traditional SMT and HPM SMT for the translation between South Asian and English languages carried out in this study. We observed HPM SMT outperform the traditional SMT for the translation between morphologically rich and poor languages (for the same dataset). Hierarchical phrase based models helps to improve translation quality between languages that vary by sentence structure. The comparison between traditional SMT and HPM SMT for the translation between Sinhala and Tamil languages also carried out in this study. However, in this case, traditional approach performs better compared to the hierarchical phrase model. Hence, hierarchical phrase based models lower the quality of languages that share similar sentence structure since built in Parser is only available for English language in Moses tool.

The challenges we faced were the lack of freely available linguistic resources and the shortage of well-developed and widely used open source frameworks. In our future work, we plan to analyze problems with existing data sets, the concern of morphology and its relation to output quality by combining those models together.

## 5 Acknowledgement

We are grateful to Treesa Anjaly Cyriac who supported in providing Malayalam resources. We are thankful to all our colleagues who supported us and to Linguists who helped in aligning POS tagsets. The authors would also like to thank the LKDomain registry for partially funding this publication.

## References

1. Chiang, David.: A hierarchical phrase-based model for statistical machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, (2007).
2. Watanabe, Taro, Hajime Tsukada, and Hideki Isozaki.: Left-to-right target generation for hierarchical phrase-based translation. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics. . (2006).
3. Chiang, David.: A hierarchical phrase-based model for statistical machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. (2005).
4. Gispert, Adrià.: Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars. Computational linguistics 36.3 (2010): 505-533. (2010).
5. Koehn, Philipp.: Edinburgh system description for the 2005 IWSLT speech translation evaluation. International Workshop on Spoken Language Translation (IWSLT) 2005. (2005).
6. Mahsa Mohaghegh, Abdolhossein Sarrafzadeh.: A Hierarchical Phrase-Based Model for English-Persian Statistical Machine Translation. 2012 International Conference on Innovations in Information Technology (IIT). (2012).
7. Jawaaid, Bushra, and Daniel Zeman.: Word-order issues in english-to-urdu statistical machine translation. *The Prague Bulletin of Mathematical Linguistics* 95: 87-106. (2011).
8. Khan, Nadeem.: English to urdu hierarchical phrase-based statistical machine translation. Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing. (2013).
9. Hermann, 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, 1–8, ACL, Morristown, NJ, USA. (2001).
10. Vasu Renganathan.: An ineractive approach to development of English to Tamil machine translation system on the web. Proceedings of INFITT-2002. (2002).
11. AUKBC Research Centre
12. Loganathan, R.: English-Tamil Machine Translation System. Master of Science by Research Thesis, Amrita Vishwa Vidyapeetham, Coimbatore. (2010).
13. Kumar, M. Anand.: Factored Statistical Machine Translation System for English to Tamil Language. *Pertanika Journal of Social Sciences & Humanities* 22.4. (2014).
14. Unnikrishnan P, Antony P J, Soman K P.: A Novel Approach for English to South Dravidian Language. (2011).

15. Sebastian, Mary Priya, K. Sheena Kurian, and G. Santhosh Kumar.: A framework of statistical machine translator from English to malayalam. Proceedings of Fourth International Conference on Information Processing, Bangalore, India. (2010).
16. R. Weerasinghe.: A statistical machine translation approach to sinhala-tamil language translation. Towards an ICT enabled Society, p. 136. (2003).
17. S. Sripirakas, A. Weerasinghe and D. L. Herath.: Statistical machine translation of systems for Sinhala-Tamil. In Advances in ICT for Emerging Regions (ICTer), 2010 International Conference. (2010).
18. Sripirakas Sakthithasan, ruvan Weerasinghe, and Dulip Lakmal Herath.: Statistical machine translation of systems for sinhala-tamil. In Advances in ICT for Emerging Regions (ICTer), 2010 International Conference on, pages 62–68. IEEE. (2010).
19. Mahendran Jeyakaran and Ruvan Weerasinghe.: A novel kernel regression based machine translation system for sinhala-tamil translation. In Proceedings of the 4th Annual UCSC Research Symposium. (2011).
20. R. Pushpananda, R. Weerasinghe and M. Niranjana.: Sinhala-Tamil Machine Translation: Towards better Translation Quality," in Australasian Language Technology Association Workshop 2014, Melbourne. (2014).
21. S. Rajpirathap, S. Sheeyam and K., C. A. Umasuthan.: Real-time direct translation system for Sinhala and Tamil languages. In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference. (2015).
22. <http://ltrc.iiit.ac.in/corpus/corpus.html>
23. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens.: Moses: Open source toolkit for statistical machine translation. (2007).
24. K. Heafield.: KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation. (2011).
25. F. Och.: Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. (2003).
26. L. Huang and D. Chiang.: Forest rescoring: Faster decoding with integrated language models. In Annual Meeting-Association For Computational Linguistics. (2007)
27. K. Papineni, S. Roukos, T. Ward and W.-J. Zhu.: BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics. (2002).
28. Nagata, Masaaki, et al.: A clustered global phrase reordering model for statistical machine translation. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics. (2006).
29. R. Pushpananda, R. Weerasinghe and M. Niranjana. Statistical Machine Translation from and into Morphologically Rich and Low Resourced Languages. in International Conference on Intelligent Text Processing and Computational Linguistics. Springer International Publishing. (2015).