Statistical Machine Learning for Transliteration: Transliterating names between Sinhala, Tamil and English

H.S. Priyadarshani, M.D.W. Rajapaksha, M.M.S.P. Ranasinghe, K. Sarveswaran, G.V. Dias
Department of Computer Science and Engineering
University of Moratuwa
Katubedda 10400, Sri Lanka
{shanika.15, wimukthi.15, sharmila.15, sarves, gihan}@cse.mrt.ac.lk

***Abstract*—In this paper, we focus on building models for transliteration of personal names between the primary languages of Sri Lanka - namely Sinhala, Tamil and English. Currently, a Rule-based system has been used to transliterate names between Sinhala and Tamil. However, we found that it fails in several cases. Further, there were no systems available to transliterate names to English. In this paper, we present a hybrid approach where we use machine learning and statistical machine translation to do the transliteration. We built a parallel trilingual corpus of personal names. Then we trained a machine learner to classify names based on the ethnicity as we found it is an influencing factor in transliteration. Then we took the transliteration as a translation problem and applied statistical machine translation to generate the most probable transliteration for personal names. The system shows very promising results compared with the existing rule-based system. It gives a BLEU score of 89 in all the test cases and produces the top BLEU score of 93.7 for Sinhala to English transliteration.**

***Index Terms*—statistical machine translation; transliteration; naive bayes; sinhala; tamil**

I. Introduction

Transliteration can be defined as the phonetic translation of names across languages [1]. However, transliteration may be based on pronunciation, spelling, or a combination of the two. It is complicated by the varied approches taken by different people. Therefore, it is a challenging task to write names in another script. One problem is that the target script may not have a way to represent the sounds in the name. Therefore, in practice names are transliterated to produce the closest pronunciation. For example, the name අනගාරික ධර්මපාල in Sinhala language can be transliterated to Tamil language as அனகாரிக்க தர்மபால and to English language as Anagarika Dharmapala which gives a close pronunciation.

Apart from pronunciation, writing names also may be influenced by other factors like numerology. For instance, people sometimes may include characters intentionally to comply with their numerology requirements. For instance, a person can write his name as Anagaarika instead of Anagarika. However, it is not very straight forward to generate the closest transliteration of names. Sinhala and Tamil are the official languages of Sri Lanka and English is the link language. All these are used widely in government documents and other varieties. The department for registration of people in Sri Lanka has a system to transliterate between Sinhala in Tamil which is used to transliterate names in national identity cards [1]. However, there are not many formal studies on transliteration among Sinhala, Tamil and English languages are found except the work by Hettige et al [2].

II. Motivation

The need for a transliteration system emerges due to several reasons. As per the policy in Sri Lanka, it is a must to represent names in all three languages. Especially, the government documents are issued mostly in all three languages and those documents contain a lot of personal names. Currently, these names are transliterated manually in government documents. Different translators find different ways to transliterate as there is no one to one mapping between Sinhala – Tamil and English characters and sounds. As an example, the characters ද, ධ, ත, ථ in Sinhala are mapped to letter த in Tamil.

Therefore phonemes which can be transcribed using Sinhala cannot be done exactly in Tamil or English and vice versa also true. One such case is that the phoneme ‘z’ is not in Sinhala or Tamil languages. As an example, how to write the name Fouza in Sinhala is somewhat confusing as some write it as ෆවුසා and some others as ෆවුzසා.

In some cases, the influence of other languages including Sanskrit, Dutch, Portuguese, Arabic can be seen in proper names and especially in personal names. As an example, names like Fernando and de Silva have a Portugese origin. In Sinhala language, those become ප්‍රනාන්දු and ද සිල්වා, which does not represent the exact phonetic mapping to English.

On the other hand, a transliteration system of this nature is also useful for Natural Language Processing (NLP) tasks such as machine translation between these languages, corpus or sentence alignment, cross-language information retrieval, information extraction and automatic lexicon acquisition.

Therefore there is a need for transliteration system which can transliterate names between Sinhala – Tamil and English Languages.

[1] https://bit.ly/2Z8TEBV

III. Related Work

There have been several research studies done on this area for different languages. Basically, transliteration is considered as a key component of translation. In this field, there are a number of models have been developed based on machine transliteration approaches such as phoneme based transliteration model, grapheme based transliteration model, hybrid transliteration model and also correspondence-based transliteration model [3].

Grapheme based transliteration model is a direct mapping of spellings or grapheme from a source language to a target language. Most of the time this is an orthogonal mapping. Channel Model and Decision Tree Model are such transliteration methods proposed using the above approach [4]. Phoneme-based transliteration model is basically built on pronunciation or the source phoneme of words in the language rather than the spellings. A Hybrid approach is one that uses both of the above for more accurate transliteration since some of these do not match with certain language specifications.

In 2007, Jiang et al [4] have proposed a method of improving the translation of named entities with the help of transliteration together with web mining. There in the transliteration model, they consider both the similarity in pronunciation and co-occurrence of bilingual contextual information of the words. A list of generated candidates is kept scoring with the help of web mining and improve the quality of the translation with better transliteration.

Later in 2014, Mathur et al [5] have reported a transliteration method for named entities in Hindi language following a Hybrid approach. There, they have used a rule-based approach for the extraction of named entities and a statistical approach in converting named entities in English to the corresponding Hindi representations. This is why it is called a hybrid approach. Further, they have applied this mechanism to their system and calculated the accuracy using precision, standard recall, and f-score. The results obtained from these experiments were compared with the results of manually transliterated named entities that was done with the help of human translators. Those results have shown better progress than the previous occasions. In 2016, Y. Lin et al [6] have reported how the transliteration methods ignore source context information and inter-dependency among entities for entity disambiguation. They bring out a novel approach to leverage state-of-the-art Entity Linking (EL) techniques to automatically correct name transliteration results, using collective inference from source contexts and additional evidence from the knowledge base.

Grundkiewicz and Heafield [7] proposed a neural machine translation based approach for transliteration in 2018 using deep attentional RNN encoder-decoder models. Mihaela and Breuel presents another approach for neural network based model for transliteration using a sequence to sequence model [8]. Their data is based on Arabic and English parallel text. Kundu et al [9] proposed an approach for transliteration based on both recurrent neural networks and convolutional sequence to sequence based neural machine translation.

When it comes to Sinhala language machine transliteration have been done for over a decade in terms of translation. In 2007, Hettige and Karunananda [2] propose a system for transliteration from English to Sinhala language. There the approach they used is based on finite-state automata. They have used a finite state automaton to develop transducers for language transliteration and the system is developed using Prolog server page and SWI-PROLOG. Those generated transducers are tested using Sinhala Chat-bot and English WordNet to obtain the expected results. They claim that handling the pronunciations of an English word is critical as one letter could have different sounds and it causes to leave some ambiguity in transliteration. The team intended to improve this system by incorporating IPA to their system as future work.

In 2010, B. Hettige et al [10] reported a methodology for English to Sinhala machine translation supported by a transliteration agent using finite-state transducers.

In 2018, Tennage et al [11] have built a system for neural machine translation supported with transliteration from Tamil to Sinhala language. This transliteration model was implemented using English as a middle language. The model has given a BLEU score of 8.36 for Tamil to Sinhala transliteration, which was a rule based approach that used character mapping without considering named entities.

In 2018, Thayaparan [12] has proposed a named entity translation model which encompassed word-embedding models to improve translation in between Sinhala and Tamil. The model was able to gain BLEU score of 35.28 for Sinhala to Tamil, and 23.75 for Tamil to Sinhala, after intergrating into existing translation system.

In 2003, Kohen et al [13] present about phrase-based machine translation and in 2007 Kohen and his team introduce Moses toolkit [14] for statistical translation. In 2009, Chinnakotla and Daman [15] talks about using Moses toolkit for transliteration, following a phrase-based SMT approach. There the words are replaced by characters and sentences are replaced by words. Again Rejwanul et al [16] follow the same approach for English-Hindi transliteration.

IV. Transliteration System

A. Our Approach

There are several transliteration systems that were built following a rule-based approach [5, 17, 18]. However, there are many cases where the rules cannot correctly handle mappings in between languages [2]. Considering these difficulties, we moved with a statistical machine translation approach. We also realised that the way names were transliterated depended on the ethnicity of the name. In Sri Lanka, the main ethnicities are Sinhalese, Tamils and Muslims. Therefore, before the transliteration of personal names we classified names based on ethnicity. This was done with a simple classification followed by the Naive Bayes algorithm. These classified names were then used to construct separate models based on both ethnicity and language. Therefore, altogether we developed 10 models for the transliteration of Sinhalese, Tamil and Muslim names in the context of Sinhala, Tamil and English languages.

In transliteration, there are several techniques which have been used in previous research studies. Direct mapping approach is one of them [19], which generates the results using parallel corpus which is given for training. It consumes minimum time than other approaches [19], however, it can transliterate only words which are present in the parallel corpus.

There are several other transliteration systems that were built basically following a rule-based approach. In this approach, different rules will be used to generate transliteration results. Rules can be created by considering the key properties of source and target languages. The rule-based approach is not widely used since it takes time, money and trained personnel to make and test the rules.

Other than the above two approaches of direct mapping and Rule-based, neural machine translation (NMT) is also a current technique which learns directly and treats the words as smallest units for transliteration. Its slower training and inference speed, ineffectiveness in dealing with rare words, and sometimes fail to translate all the words in the source sentence. NMT lacks robustness in translating rare words [20] and it needs large parallel data sets to train the model to obtain better results than SMT [11].

Considering those difficulties, we used a statistical machine translation (SMT) approach. SMT is a language-independent and data-oriented approach to transliterate text from a source language to a target language. SMT has high accuracy results and time efficient than rule-based approach [19]. SMT transliterates not only Sinhala origin names but other names as well. When corpus size is small, SMT performs better than the NMT according to results obtained by Tennage et al. [21]

Our transliteration system development involves the following steps:

- Preprocessing the data in Sinhala - Tamil parallel name corpus:
  - cleaning the data to remove duplicates, correct spelling, correct Unicode errors and break names to labels
- Training and tuning models for Sinhala to Tamil and Tamil to Sinhala transliteration using Moses decoder [2]. Then testing the models, measuring the BLEU score and identifying the issues with transliteration model.
- Manually classifying a part of Sinhala Tamil parallel name corpus based on the ethnicity of the names.
- Creating a model for ethnicity-based classification of names using the manually classified data using the Naive Bayes algorithm.
- Classifying the rest of the data in the parallel corpus with the built classification models, and manually validating the accuracy.
- Retraining models for Sinhala to Tamil and Tamil to Sinhala transliteration based on ethnicity and building six models representing each transliteration with Moses decoder. Then tuning and testing each model for the BLEU score.
- Conducting a survey to collect proper names from all three languages.
- Scraping web and collecting a list of English proper names. Then transliterating them to Sinhala language using a rule-based approach and manually correcting the result and building an English Sinhala parallel name corpus.
- Building models for Sinhala to English, and English to Sinhala Transliteration, with the data from English Sinhala parallel corpus, and testing for the accuracy of the model.
- Using the developed Sinhala to English Transliteration model, transliterating a set of Sinhala names from the Sinhala English name corpus, to English language. Then manually validating the names and creating an English Tamil parallel corpus.
- Building the models for English to Tamil and Tamil to English transliteration with Moses decoder, using the English Tamil parallel data. Then tuning and testing the models for the accuracy.

B. Dataset

A parallel name corpus of 100,000 Sinhala and Tamil was obtained from a government department. However, some of these names, especially Sinhala names had multiple tokens in their names. For instance, the name Chathuri Ishaka Harshani has three tokens all belonging to one personal name and also mapped to corresponding Tamil labels in the corpus.

However, no publically available Sinhala-English or Tamil-English transliterator was found. Therefore, we tried a survey approach to collect names from university students and the general public through a survey and we ended up collecting 2000 names which were insufficient for us to train the system. Then we crawled the web to find proper names in English, mostly from the websites with examination results published. Finally, we collected around 80000 proper names in English. These names were then transliterated to Sinhala language with a rule-based approach followed by H.M. Weerasingha [17] which had a BLEU score of 80.03%. However, there were a lot of issues with the transliterated results, some of which are shown in Table 1.

It is not that these transliterations are incorrect, but when it comes to proper names in Sri Lankan context the name in the last column is preferred or widely used than the transliterated output in the second one as in Table 1. Therefore, all these transliterations were checked for errors and corrected manually. But still, there are some ambiguities in names, especially gender-wise. As an example, the name Maneesha can be transliterated either as මනීෂා or මනීෂ, a depending on whether the name is female or male.

[2] https://github.com/moses-smt/mosesdecoder

TABLE I
Examples for confusing results from English to Sinhala rule-based Transliteration

Name in English	Transliterated result in Sinhala	The expected result in Sinhala
Sandakelum	සන්දකෙළුම (sʌndʌkelʊmʌ)	සදකැලුම් (sʌndʌkælʊm)
Imasha	ඉමෂ (imʌʃʌ)	ඉමාෂා/ඉමාෂ (imɑːʃɑː / imɑːʃʌ)
Menike	මෙනික (Menikʌ)	මැණිකේ (Mænikeː)
Yogaraj	යොගරජ් (jogʌrʌʤ ʥ)	යෝගරාජ් (joːgʌrɑːʤ ʥ)
Margret	මර්ග්‍රෙට් (mʌrgreʈ)	මාග්‍රට් (mɑːgrʌʈ)

C. Model Creation

In each model creation process, between pairs of two languages from Sinhala, Tamil and English, the Moses-decoder was fed with the parallel corpus of personal names in source and target languages. The corpus contained names in random order and it was also partitioned to a ratio of 5:2:1 for training, tuning and testing respectively. We convered words into a sequence of characters, i.e., we used character segmentation rather than word segmentation.

Before training the transliteration model, a language model was built with the target language using KenLM. It was to find the most widely used or preferred from the number of outputs generated by the SMT. In this case, a three-gram language model is generated and also binarised with KenLM[3] to achieve faster loading. We used Giza$++$[4], which is the default tool in Moses, to build the translation models, in our case the transliteration models. Once the extraction of terms, scoring and lexicalized reordering tables creations are done, the final Moses configuration file is taken as the output of the training phase. Then each model is tuned for better results and tested for the BLEU (Bilingual Evaluation Understudy) score.

D. Sinhala-Tamil Transliteration Model

The first model we created was to transliterate between Sinhala and Tamil languages. As stated above it was done with a parallel corpus of 100,000 entries of full names. Though the model gave good results, still there were some confusing cases and some such examples are displayed in Table 2 and Table 3.

In these cases, also, it is not that the character mapping in transliteration is incorrect, but the possibility of having such names is rare. However, when analysing further we found that this irregularities arose may be due to the ethnicity of those names belongs to. In Sri Lanka, the way names are written in a language depends on their ethnicity. At the same time, the mapping in between the characters in Sinhala, Tamil and English names are not one to one. Therefore, we had to address this issue of diversity by classifying names based on these irregularities in ethnicity.

TABLE II
Examples for confusing results from Sinhala to Tamil Transliteration

Input name in Sinhala	Transliterated result in Tamil	The expected result in Tamil
රාමනායකලාගේ (rɑːmʌnɑːjʌkʌlɑːge)	ராமநாயக்கலாகே (ɾ̪aːman̪aːjakal̪aːkeː)	ராமனாயகலாகே (ɾ̪aːman̪aːjakal̪aːkeː)
වේලුසාමි (ʋeːlʊsɑːmi)	வேலுசாமி (ʋeːl̪usaːmi)	வேலுச்சாமி (ʋeːl̪ussaːmi)
සෙල්වරාජා (Selʋʌrɑːʤ ʥɑː)	செல்வராஜா (sel̪ʋaɾ̪aːd͡ʒaː)	செல்வராசா (sel̪ʋaɾ̪aːsaː)
අවුසෙෆ් (aʊʊsef)	அவுசெப் (aʋusep)	ஓவுசெப் (oːʋusep)
මැරික්කාර් (mærikkʌr)	மெரிக்கார் (meɾ̪ikkaːɾ̪)	மரிக்கார் (maɾ̪ikkaːɾ̪)

TABLE III
Examples for confusing results from Tamil to Sinhala Transliteration

Input name in Tamil	Transliterated result in Sinhala	The expected result in Sinhala
அபிலாஷா	අපිලාෂා (apilɑːʃɑː)	අභිලාෂා (abilɑːʃɑː)
கங்கா	කංකා (kʌŋkɑː)	ගංගා (gʌŋgɑː)
சித்திரவேல்	සිත්තිරවේල් (sittirʌʋeːl)	සිද්‍රවේල් (sidrʌʋeːl)
ஜெப்றின்	ජෙබ්රින් (ʤ ʥbrin)	ජෙෆ්රින් (ʥfrin)
வகாப்தீன்	වහාප්දීන් (ʋʌɦɑːpdiːn)	වහාබ්දීන් (ʋʌɦɑːbdiːn)

E. Classification of personal names based on ethnicity

In most cases of Sri Lankan context, a name could reveal the person's ethnic group and the names have their own specifications based on that origin. In this case, we observed that there is a significant pattern of transliterating a name based on ethnicity. Therefore, the names were first classified according to their ethnicity (Sinhalese, Tamil or Muslim) before transliteration.

Then we built a machine learning model to classify the personal names into Sinhalese, Muslim and Tamil names. In making the training data set for the classification model, we manually classified around 30,000 full names into each category. Next, we split the dataset into training and validation sets so that we can train and test the classifier. Also, we encoded our target column so that it can be used in machine learning models.

As features, raw text data was transformed into feature vectors and new features were created using the existing dataset. We implemented TF[5]-IDF[6] (Term Frequency - Inverse Document Frequency) vectors as features in order to get relevant features. TF-IDF score represents the importance of terms appears in the entire corpus. We considered two feature vectors as N-gram level TF-IDF and character level TF-IDF. N-gram level TF-IDF vector represents TF-IDF scores

[3] https://kheafield.com/code/kenlm/

[4] https://github.com/moses-smt/giza-pp

[5] TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

[6] IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

of N terms together and character level TF-IDF represents the scores of character level N-grams in the corpus. After analyzing the accuracy of models using both feature vectors, Character level N-gram TF-IDF feature give better results than the other features.

Finally, we implemented a Naive Bayes model using Sklearn[7] implementation with different features. Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Then, we trained the classifier using the training data set and validated using the rest of the data. This model gave a 96.1% accuracy on ethnicity-based classification. Sinhala Tamil Transliteration Models based on ethnicity Using the classified names from the classification model, we trained three separate models based on ethnicity as Sinhalese, Tamil and Muslim, by feeding the Moses decoder with relevant data for each language pair for each ethnic categories using the same way as we followed earlier. The transliterations were done both back and forth between Sinhala and Tamil languages. This approach gave better results than the previous time and most of the confusions in transliteration mapping were resolved.

F. Sinhala English Transliteration Models

As described previously, using the trilingual corpus we created, two other models were built for back and forth transliteration in between Sinhala and English languages. In this case, the models gave a good accuracy even without classification on ethnicity. Therefore, for this direction, we considered names belonging to all ethnic groups together as it is an overhead to create more models.

G. Tamil English Transliteration Models

The rule-based approach was giving poor results in the transliteration of names from English to Tamil. Therefore, to generate an English Tamil parallel corpus with personal names, we used the models trained by Moses for English to Sinhala transliteration and Sinhala to Tamil transliteration. They were giving better results than the Rule-based approach and still we had to correct them manually to create a Tamil and English parallel corpus with no spelling mistakes.

V. Evaluation And Discussion

Sinhala to Tamil and Tamil to Sinhala transliterations models without ethnicity-based classification were based on a corpus of all types of personal names found in Sri Lanka. There the model for Sinhala to Tamil transliteration gave a BLEU score of 82.33% while the model for Tamil to Sinhala transliteration presented a BLEU score of 80.02%. There were several issues with these transliterated results caused by the variations of the originality of these proper names. As an example, mostly in the Tamil language, the letter 'க்' represents the [h] phoneme in personal names. However, when it comes to Muslim names they are used to write with the 'ஹ்' character, which is a Grantha consonant but not used widely in the names of Tamils. In some cases, some letters are missing in source or target language. As an example, there is no separate phoneme for 'ෆ' [f] in Tamil language. Then the name ෆාතිමා (Fathima) is written as பாத்திமா in Tamil where character 'ப' represents the phone for [p] as well. The name පාරමී is also written using the same character, as பாரமீ but representing the phoneme [p]. Again all the phonemes 'ක', 'ඛ', 'ග', 'ඝ' in Sinhala language is mapped to single 'க' in Tamil language in general use. In all these cases the choice was mainly based on the ethnic group. This is one example and there are many more such many to one or one to many mappings that confuses the system when all types of names are considered together. But, this confusion gets decreased, once when we consider the ethnicity behind the origin of these names. BLEU score quality metric increased with the application of the classification model. The system shows the BLEU score greater than 89% for all pairs of transliteration. For the transliteration from Sinhala to English, a BLEU score of 93.7% was obtained for the backward transliteration the score was 92.37%. All the BLEU Scores used in this paper are in BLEU-4 metric.

TABLE IV
BLEU Scores in ethnicity-based models

Ethnicity	Source Language	Target Language	BLEU Score (%)
Sinhalese	Sinhala	Tamil	89.15
Sinhalese	Tamil	Sinhala	91.47
Tamil	Sinhala	Tamil	93.62
Tamil	Tamil	Sinhala	91.29
Muslim	Sinhala	Tamil	89.35
Muslim	Tamil	Sinhala	89.61

TABLE V
BLEU Scores in Tamil-English and Sinhala-English models

Source Language	Target Language	BLEU Score (%)
Sinhala	English	93.70
English	Sinhala	92.37
Tamil	English	86.11
English	Tamil	91.74

VI. Conclusion

In this paper, we have presented a statistical machine translation approach to transliterate personal names in Sri Lankan context using Moses SMT toolkit for Sinhala, Tamil and English languages. We have improved the results further by using the ethnic origin of a given name, whether Sinhalese, Tamil or Muslim. We developed a classification model to classify names before feeding to Moses for transliteration. Our system shows a BLEU score of more than 89% for all the language pairs of consideration.

VII. Future Work

The personal name transliteration module is just a part of our named entity translation project. Therefore we will extend this to cover location names, organizational names and designations. In these cases, just transliteration would not

[7] https://github.com/scikit-learn/scikit-learn

be enough. Therefore, we will also use other techniques like terminology integration to improve quality.

References

[1] N. Chen, X. Duan, M. Zhang, R.E. Banchs , H. Li, “Whitepaper on NEWS 2018 Shared Task on Machine Transliteration”

[2] B. Hettige and A. S. Karunananda, ”Transliteration system for English to Sinhala machine translation,” 2007 International Conference on Industrial and Information Systems, Peradeniya, 2007, pp. 209-214.

[3] O. Jong-Hoon, C. Key-sun, I. Hitoshi, “A comparison of Different Machine Transliteration models”, Journal of Artificial Intelligence Research, pp 119- 151, 2007.

[4] L. Jiang, M. Zhou, L.F. Chien, C. Niu, “Named Entity Translation with Web Mining and Transliteration”, The International Joint Conference on Artificial Intelligence, Inc. (pp. 1629-1634). Hyderabad: Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2007

[5] S. Mathur, V. P. Saxena, ”Hybrid approach to English-Hindi name entity transliteration,” 2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science, Bhopal, 2014, pp. 1-5

[6] Y. Lin, X. Pan, A. Deri, H. Ji, K. Knight, “Leveraging Entity Linking and Related Language Projection to Improve Name Transliteration”, 2016

[7] R. Grundkiewicz, K. Heafield “Neural Machine Translation Techniques for Named Entity Transliteration”, Proceedings of the Seventh Named Entities Workshop, July 2018

[8] M. Rosca, T. Breuel “Sequence-to-sequence neural network models for transliteration”,2016

[9] S. Kundu, S. Paul and Santanu Pal, “A Deep Learning-Based Approach to Transliteration”,2018

[10] B. Hettige and S. K. Asoka, ”An evaluation methodology for English to Sinhala machine translation,” 2010 Fifth International Conference on Information and Automation for Sustainability, Colombo, 2010, pp. 31-36.

[11] Tennage, P., Herath, A., Thilakarathne, M., Sandaruwan, P. (2018). Transliteration and Byte Pair Encoding to Improve Tamil to Sinhala Neural Machine Translation. Moratuwa Engineering Research Conference (MERCon). Moratuwa, Sri Lanka: IEEE.

[12] M. Thayaparan, “Translation of Named Entities Between Sinhala and Tamil for Official Government Documents”, M.S. thesis, Dept. of Comp. Science and Eng, Univ. Moratuwa, Sri Lanka, 2018.

[13] P. Koehn, F. J. Och, D. Marcu. 2003. “Statistical phrase-based translation”, Proc. of HLTNAACL 2003, Edmonton, Canada, pp. 48-54

[14] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. “Moses: open-source toolkit for statistical machine translation”, Proc. of ACL, Prague, Czech Republic, pp. 177- 180.

[15] M. K. Chinnakotla, O. P. Damani. 2009. “Experiences with English-Hindi, English-Tamil and English-Kannada transliteration tasks at NEWS 2009”, In Proc. ACL/IJCNLP Named Entities Workshop Shared Task.

[16] R. Haque, S. Dandapat, A. K. Srivastava, S. K. Naskar, A. Way, “English-Hindi transliteration using context in formed PB-SMT”, In Proc. ACL/IJCNLP Named Entities Workshop Shared Task, 2009.

[17] H. M. Weerasinghe, “Transliteration of Names from English to Sinhala”, M.S. thesis, Dept. of Comp. Science and Eng, Univ. Moratuwa, Sri Lanka, 2006.

[18] S.C. Fernando, “Inexact matching of proper names in Sinhala”, M.S. thesis, Dept. of Comp. Science and Eng, Univ. Moratuwa, Sri Lanka, 2007 .

[19] V. Kaur, A. K. Sarao, J. Singh, “A Review on Hindi to English Transliteration System for Proper Nouns Using Hybrid Approach”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 5, September 2014.

[20] Y. Wu, M. Schuster, Z. Chen, Q. V Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. “Google’s neural machine translation system: Bridging the gap between human and machine translation”. arXiv preprint arXiv:1609.08144, 2016.

[21] P. Tennage, P. Sandaruwan, M. Thilakarathne, A. Herath, S. Ranathunga, ”Neural Machine Translation for Sinhala and Tamil Languages”, in International Conference on Asian Language Processing, 2017.