A Novel Hybridized Strategy for Machine Translation of Indian Languages

Santhanavijayan A[1], Naresh Kumar D[2*], Gerard Deepak[1]
Department of Computer Science & Engineering
[*]Department of Mechanical Engineering
National Institute of Technology, Tiruchirappalli, India.
[*]d.nareshkumar1040@gmail.com

Abstract. Only a few translation systems exist for translations between an Aryan language and a Dravidian language and moreover, the existing do not perform quite efficiently. In this paper, the cognitive gap between two Indian languages namely Hindi and Malayalam is bridged. This paper proposes a Hindi to Malayalam Machine Translation system using Hybridized Strategies such as Phrase-Based Translation, Word Alignment and a Language Model with the emphasis on Transition Probability computation. Phrase-based translation breaks down a sentence into phrases and translating each phrase independently This technique is applied on a Hindi-Malayalam parallel corpus. The proposed model interoperates between Hindi and Malayalam Languages. Although standard Natural Language Computing Techniques are encompassed, the arrangement of the techniques to suit a pair of Indian Languages that are semantically incompliant and achieving a high accuracy is definitely a challenge and is achieved in this paper. The proposed hybridized approach outperforms all the other existing strategies and yields an average precision of 90.725% with a low word error rate of 9.125.

Keywords: Language Model, Machine Translation, Phrase-based translation, Word Alignment

1 Introduction

In simple terms, machine translation is automatic translation carried out by a computer and not by humans This can be helpful in translating languages which a particular human is not familiar with. Machine Translation can be trained to become more accurate i.e. they can learn from manual translations. Two most MT (Machine Translation) engines used are RBMT (Rule-Based Machine Translation) and SMT (Statistical Machine Translation). These differ in the way that they process and analyze contents. They can also be combined within the same system which gives rise to hybrid MT. RBMT is based on linguistic information. This engine translates the source text into the desired language by covering all the main semantic, morphological and syntactic regularities of each language. RBMT uses linguistic rules to break down the content. It produces more predictable outputs in terms of terminology and grammar through the use of customizable terminology lists to fine tune the engine. They do not need bilingual corpus to create the translation system.

SMT is another machine translation engine that is generated on the basis of statistical models. The parameters of these statistical models are derived from the analysis of bilingual text corpora. This statistical approach is much different and contrasts with the above mentioned RBMT. Information theory gives the foundational idea for SMT. The given input document is translated by the usage of Probability distribution. This probability distribution can be achieved by many procedures like that of Bayes theorem. In this paper a novel method based on inter-language word alignment between a pair of Indian languages has been proposed as a specific case we have chosen Hindi and Malayalam. A phrase-based translation model is designed for the same. The phrases follow a 4-gram model. The system takes as input a source text file of Hindi sentences and gives an output text file of its corresponding Malayalam translation.

The remaining paper organization is as follows. Section 2 presents related work. Section 3 presents the proposed system architecture. Section 4 depicts the results and performance evaluation. Section 5 concludes the paper.

2 Related Work

Aditya Kaustav et al.,[1] have proposed a constructive method to translate English language to Odia language besides giving a clear definition of various translation methods widely used. Sindhu et al., [2] give a comparative study of all the machine translation approaches existing for Indian languages. It further delves into the advantages and disadvantages of all the methods mentioned and the various challenges suffered by such systems. Pasindu et al., [3] have proposed a Neural Machine Translation for Sinhala and Tamil languages which performs really well but definitely is much more complex. Raihan et al., [4] give an efficient method for aligning sentences from English-Bengali parallel corpora.IT leverages the lexical information of the language pair to optimize the system. Issues in translating from a comparable corpus are suggested in the paper written by Tholpadi et al., [5] This work uses an auxiliary language to optimize the model along with providing new datasets. Jadoon et al., [6] have presented an analysis on these with a special focus in Indian languages taking as example eight prominent Indian languages. Combining various MT techniques give rise to better translation. One such method for Hindi-English language pair has been proposed by Omkar et al., [7] This paper combines SMT, EBMT, and RBMT. Ajit Kumar; Nitin Bansal et al [8] gives a survey of all the machine translation techniques available for two similar regional Indian languages namely Punjabi and Urdu. The paper also mentions various methods to evaluate such MT systems. Archana G P et al., forays into this particular area of an MT system. [9] This particular paper lists all the issues and challenges one encounter in creating a viable parallel corpus. However, incorporation of semantic techniques as in [10-24] would definitely yield a much more promising technique.

3 System Architecture

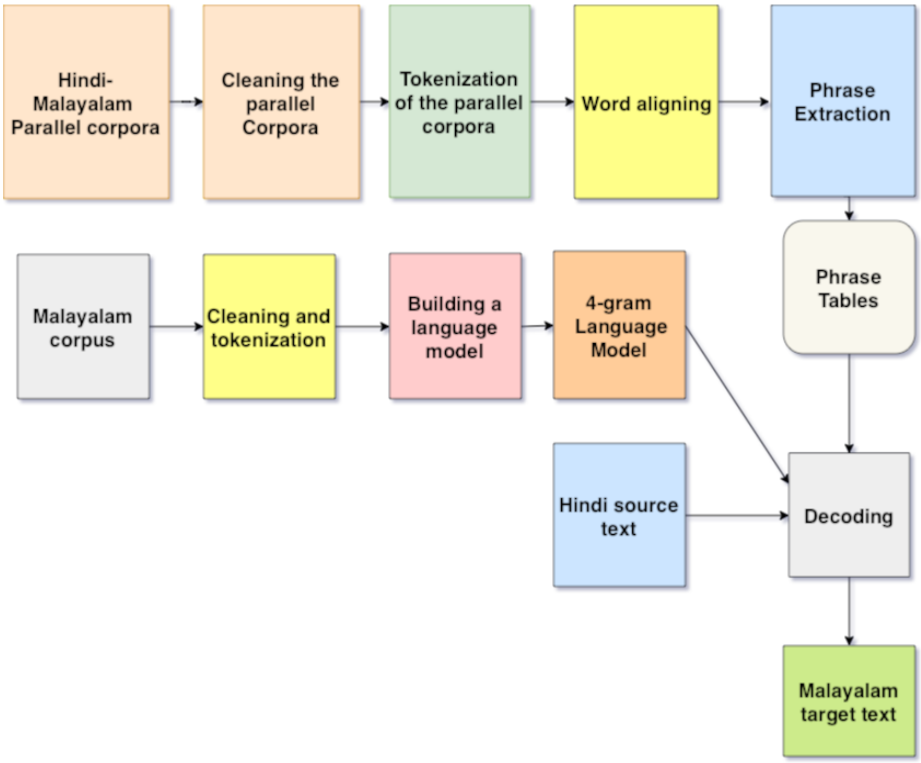

Hindi-Malayalam Parallel corpora
Cleaning the parallel Corpora
Tokenization of the parallel corpora
Word aligning
Phrase Extraction
Phrase Tables
Malayalam corpus
Cleaning and tokenization
Building a language model
4-gram Language Model
Hindi source text
Decoding
Malayalam target text

Figure 1: Proposed System Architecture

The proposed system architecture is given in Figure 1 which takes in the Hindi-Malayalam Parallel Corpus as the principal input. The parallel corpus contains text and their translation in parallel form. It is entitled to include bilingual texts as well as texts that fit under translation. A parallel corpus mainly consists of two parts; a text is taken from a source language and its text taken from a target language. A small portion of TLDIR agriculture data is imbibed along with a portion of processed film subtitles. The parallel corpus is initially cleaned and further tokenized. The TLIDR dataset was already sentence aligned. For Tokenization, Indic tokenizer specifically designed for Indian scripts has been made use. The input corpus for training is fed into the Indic tokenizer package and the tokenized output will be available in the output file. Constructing parallel corpus from movie subtitles require more work. All subtitles are available in .srt format. The first step in processing subtitles is to convert this into text files. The subtitles will contain IDs and time frames for each sentence. The next step is to remove these. Sometimes the same sentence can appear more than once. Removal of duplicate sentences is an essential step. The final step is converting the subtitles into Unicode UTF-8 format.

The word alignment is an important phase for machine translation which a series of sequences are carried out. Different languages have different word order which has to be learned from data. Considering a corpus of (h,m) pairs, such that h={h_i} represents the Hindi source and m={mj} denotes the Malayalam target, the word alignment between {h,m} has to be predicted. Bayes Rule forms the basis of word alignment models. Word alignments are represented using a word alignment matrix as in Figure 3. In this paper, we use the GIZA++ tool for constructing word alignments. Alignment training in GIZA++ is done using the EM algorithm. The two most predominant steps of the EM algorithm are the initialization of model parameters and assignment of missing data probabilities and is repeated done until convergence. Further to this, Phase Extraction is done. Phrase extraction refers to extracting phrase-based lexicons. A phrase-based lexicon pairs strings in one language with strings in another language. After running the GIZA++ tool, the two files of word alignments in both directions are obtained, from which the intersection of two alignments as an initial point is considered. The phrases are further grown by adding consistent phrase pairs. In this work, the SRILM has been incorporated for a 4-gram Malayalam language model creation. The Language models contribute to the fluency of the target language., that follow the N-Gram Markov Assumptions. The transition probability focuses on the calculation of occurrence of target phrase for given source phrase in both directions. Finally, the decoding is done which involves inputting a Hindi language file and gives the corresponding Malayalam translation as the output file. The translation is done based on hypothesis recombination. Consider we have a translation model. Decoding is the process of finding the best translation with the highest probability. There are mainly two types of hypothesis-based decoding namely hypothesis expansion and hypothesis recombination. In hypothesis recombination, different hypothesis paths lead to the same partial translation. We combine the paths by either eliminating a weaker path or by giving a pointer from the weaker path. It is not necessary that the recombined hypothesis should be matched completely.

4. Results and Performance Evaluation

The implementation is carried out using Python 3.6 and Ubuntu as the preferred operating system. A chronological output screenshot of Transition Probability computation is depicted in Figure 2. Figure 3 depicts the small portion of the depiction of Word Alignment. BLEU, Word Error Rate (WER), Precision, Recall, and Accuracy are used as potential metrics for evaluation of the proposed system. BLEU is one of the most used automated metrics in machine translation. BLEU matches against a set of references translations for a greater variety of expressions. Higher the BLEU score better the translation. BLEU computation is depicted in Eq. (1), Eq. (2) and q. (3). WER is based on edit distance and is depicted in Eq. (4). Edit distance gives the minimum number of transformations needed to convert a sentence into a reference sentence. Precision and Recall is represented in Eq. (5) and Eq. (6) respectively.

$$Unigram\ precision\ \ P = \frac{\mathrm{m}}{\mathrm{Wt}} \tag{1}$$

$$Brevity\ penalty\ p = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases} \qquad (2)$$

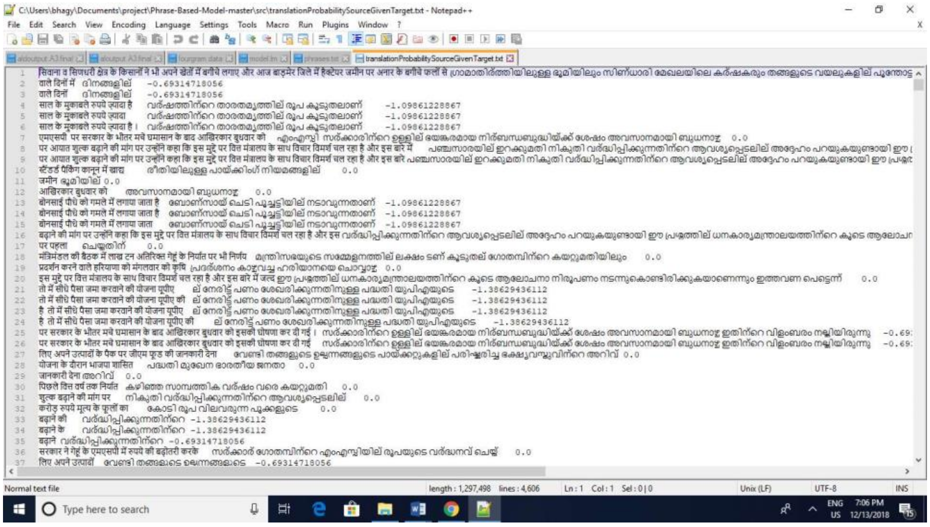
$$BLEU\ = \mathrm{p}.e^{\sum_{n=1} {}^{N}}\left(\frac{1}{N} * logPn\right) \qquad (3)$$

BLEU Calculation

$$WER = \frac{Substitution+Insertions+Deletions}{reference\ length} \qquad (4)$$

$$Precision = \frac{Correct\ number\ of\ words}{length\ of\ output\ words} \qquad (5)$$

$$Recall = \frac{Correct\ number\ of\ words}{length\ of\ reference\ words} \qquad (6)$$


C:\Users\bhagy\Documents\project\Phrase-Based-Model-master\src\translationProbabilitySourceGivenTarget.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
aldoutput A3.final
aloutput A3.final
fourgram data
model.lm
phrases.txt
translationProbabilitySourceGivenTarget.txt
Normal text file
length : 1,297,498 lines : 4,606
Ln : 1 Col : 1 Sel : 0 | 0
Unix (LF)
UTF-8
INS
Type here to search
ENG
US
7:06 PM
12/13/2018

```
सिवाना व सिणधरी क्षेत्र के किसानों ने भी अपने खेतों में बगीचे लगाए और आज बाड़मेर जिले में हैक्टेयर जमीन पर अनार के बगीचे फलों से ഗ്രാമാതിർത്തിയിലുള്ള ഭൂമിയിലും സിണ്ധാരി മേഖലയിലെ കർഷകരും തങ്ങളുടെ വയലുകളില് പൂന്തോട
वाले दिनों में   ദിനങ്ങളില്    -0.69314718056
वाले दिनों    ദിനങ്ങളില്    -0.69314718056
साल के मुकाबले रुपये ज्यादा है    വർഷത്തിന്റെ താരതമ്യത്തില് രൂപ കൂടുതലാണ്      -1.09861228867
साल के मुकाबले रुपये ज्यादा     വർഷത്തിന്റെ താരതമ്യത്തില് രൂപ കൂടുതലാണ്      -1.09861228867
साल के मुकाबले रुपये ज्यादा है।  വർഷത്തിന്റെ താരതമ്യത്തില് രൂപ കൂടുതലാണ്      -1.09861228867
एमएसपी पर सरकार के भीतर मचे घमासान के बाद आखिरकार बुधवार को     എംഎസ്പി സർക്കാരിന്റെ ഉള്ളില് ഭയങ്കരമായ നിർബന്ധബുദ്ധിയ്ക്ക് ശേഷം അവസാനമായി ബുധനാഴ്ച   0.0
पर आयात शुल्क बढ़ाने की मांग पर उन्होंने कहा कि इस मुद्दे पर वित्त मंत्रालय के साथ विचार विमर्श चल रहा है और इस बारे में      പഞ്ചസാരയില് ഇറക്കുമതി നികുതി വർദ്ധിപ്പിക്കുന്നതിന്റെ ആവശ്യപ്പെടലില് അദ്ദേഹം പറയുകയുണ്ടായി ഈ (
पर आयात शुल्क बढ़ाने की मांग पर उन्होंने कहा कि इस मुद्दे पर वित्त मंत्रालय के साथ विचार विमर्श चल रहा है और इस बारे പഞ്ചസാരയില് ഇറക്കുമതി നികുതി വർദ്ധിപ്പിക്കുന്നതിന്റെ ആവശ്യപ്പെടലില് അദ്ദേഹം പറയുകയുണ്ടായി ഈ പ്രശ്ന
स्टैंडर्ड पैकिंग कानून में खाद्य       രീതിയിലുള്ള പായ്ക്കിംഗ് നിയമങ്ങളില്      0.0
जमीन ഭൂമിയില് 0.0
आखिरकार बुधवार को       അവസാനമായി ബുധനാഴ്ച      0.0
बोनसाई पौधे को गमले में लगाया जाता है    ബോൺസായ് ചെടി പൂച്ചട്ടിയില് നടാവുന്നതാണ്  -1.09861228867
बोनसाई पौधे को गमले में लगाया जाता है    ബോൺസായ് ചെടി പൂച്ചട്ടിയില് നടാവുന്നതാണ്  -1.09861228867
बोनसाई पौधे को गमले में लगाया जाता      ബോൺസായ് ചെടി പൂച്ചട്ടിയില് നടാവുന്നതാണ്  -1.09861228867
बढ़ाने की मांग पर उन्होंने कहा कि इस मुद्दे पर वित्त मंत्रालय के साथ विचार विमर्श चल रहा है और इस വർദ്ധിപ്പിക്കുന്നതിന്റെ ആവശ്യപ്പെടലില് അദ്ദേഹം പറയുകയുണ്ടായി ഈ പ്രശ്നത്തില് ധനകാര്യമന്ത്രാലയത്തിന്റെ കൂടെ ആലോചന
पर पहला      ചെയ്തതിന്      0.0
मंत्रिमंडल की बैठक में लाख टन अतिरिक्त गेहूं के निर्यात पर भी निर्णय    മന്ത്രിസഭയുടെ സമ്മേളനത്തില് ലക്ഷം ടണ് കൂടുതല് ഗോതമ്പിന്റെ കയറ്റുമതിയിലും      0.0
प्रदर्शन करने वाले हरियाणा को मंगलवार को कृषि  പ്രദർശനം കാഴ്ചവച്ച ഹരിയാനയെ ചൊവ്വാഴ്ച  0.0
इस मुद्दे पर वित्त मंत्रालय के साथ विचार विमर्श चल रहा है और इस बारे में जल्द ഈ പ്രശ്നത്തില് ധനകാര്യമന്ത്രാലയത്തിന്റെ കൂടെ ആലോചനാ നിരൂപണം നടന്നുകൊണ്ടിരിക്കുകയാണെന്നും ഇത്തവണ പെട്ടെന്ന്      0.0
तो में सीधे पैसा जमा करवाने की योजना यूपीए      ല് നേരിട്ട് പണം ശേഖരിക്കുന്നതിനുള്ള പദ്ധതി യുപിഎയുടെ      -1.38629436112
तो में सीधे पैसा जमा करवाने की योजना यूपीए की   ല് നേരിട്ട് പണം ശേഖരിക്കുന്നതിനുള്ള പദ്ധതി യുപിഎയുടെ      -1.38629436112
है तो में सीधे पैसा जमा करवाने की योजना यूपीए   ല് നേരിട്ട് പണം ശേഖരിക്കുന്നതിനുള്ള പദ്ധതി യുപിഎയുടെ      -1.38629436112
है तो में सीधे पैसा जमा करवाने की योजना यूपीए की       ല് നേരിട്ട് പണം ശേഖരിക്കുന്നതിനുള്ള പദ്ധതി യുപിഎയുടെ      -1.38629436112
पर सरकार के भीतर मचे घमासान के बाद आखिरकार बुधवार को इसकी घोषणा कर दी गई।  സർക്കാരിന്റെ ഉള്ളില് ഭയങ്കരമായ നിർബന്ധബുദ്ധിയ്ക്ക് ശേഷം അവസാനമായി ബുധനാഴ്ച ഇതിന്റെ വിളംബരം നല്കിയിരുന്നു    -0.69
पर सरकार के भीतर मचे घमासान के बाद आखिरकार बुधवार को इसकी घोषणा कर दी गई    സർക്കാരിന്റെ ഉള്ളില് ഭയങ്കരമായ നിർബന്ധബുദ്ധിയ്ക്ക് ശേഷം അവസാനമായി ബുധനാഴ്ച ഇതിന്റെ വിളംബരം നല്കിയിരുന്നു    -0.69
लिए अपने उत्पादों के पैक पर जीएम फूड की जानकारी देना      വേണ്ടി തങ്ങളുടെ ഉല്പന്നങ്ങളുടെ പായ്ക്കറ്റുകളില് പരിഷ്കരിച്ച ഭക്ഷ്യവസ്തുവിന്റെ അറിവ് 0.0
योजना के दौरान भाजपा शासित     പദ്ധതി മുഖേന ഭാരതീയ ജനതാ    0.0
जानकारी देना അറിവ്    0.0
पिछले वित्त वर्ष तक निर्यात   കഴിഞ്ഞ സാമ്പത്തിക വർഷം വരെ കയറ്റുമതി     0.0
शुल्क बढ़ाने की मांग पर     നികുതി വർദ്ധിപ്പിക്കുന്നതിന്റെ ആവശ്യപ്പെടലില്      0.0
करोड़ रुपये मूल्य के फूलों का      കോടി രൂപ വിലവരുന്ന പൂക്കളുടെ      0.0
बढ़ाने की     വർദ്ധിപ്പിക്കുന്നതിന്റെ -1.38629436112
बढ़ाने के     വർദ്ധിപ്പിക്കുന്നതിന്റെ -1.38629436112
बढ़ाने वर्द्धिപ്പിക്കുന്നതിന്റെ -0.69314718056
सरकार ने गेहूं के एमएसपी में रुपये की बढ़ोतरी करके     സർക്കാർ ഗോതമ്പിന്റെ എംഎസ്പിയില് രൂപയുടെ വർദ്ധനവ് ചെയ്ത്      0.0
लिए अपने उत्पादों    വേണ്ടി തങ്ങളുടെ ഉല്പന്നങ്ങളുടെ   -0.69314718056
```

Figure 2: Obtained Translation Probabilities

Table 1: Performance Evaluation for the Proposed Approach using two different Corpora

Metrics	**Corpus 1**	**Corpus 2**
BLEU	70.61	68.71
WER	8.41	9.84
Precision%	91.45	90.0
Recall%	81.5	82.34

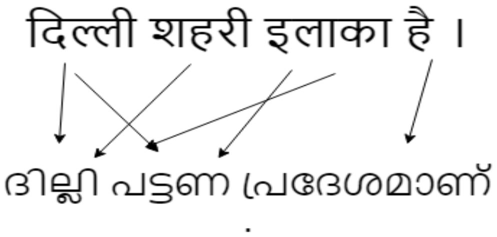


Figure 3: Word alignment

The comparatively high score of the BLEU metric is because of the in-domain nature of the data set used. Both the training set and the test set is particularly focused on agriculture and a single movie subtitle. The high value of BLEU is definitely a good indication of high-quality translations achieved. From Table 1, it is clearly evident that the proposed approach yields a high BLEU, Precision and Recall Scores with a definitely lower WER score. For corpus 1, a BLEU of 70.61 has been achieved with a precision of 91.45 and a recall of 81.5. The WER for corpus 1 is 8.41. However, for Corpus 2, the BLEU score of 68.71 is achieved with a Precision of 90 and a Recall of 82.34. The WER score for corpus 2 is 9.84. The reason for achieving a high value of BLEU score is due to the fact that word alignments are performed effectively at first. Furthermore, the phrase extraction phase ensures the addition of correctness. The 4-Gram Language Model that has been incorporated increases the density of lexical-grammatical knowledge, which ensures that a higher value of Precision and Recall are obtained. The transition probability derivation makes this technique quite feasible, efficient and lowers the WER score. However, each of these phases that have been mentioned is performed with a high degree of care with the best tools like SRILM, GIZA++ and the Indic Tokenizer for tokenization. This is the best in class performance measures that have been achieved for Hindi to Malayalam Translation, in the Indian Context.

5 Conclusions

A novel strategic method for machine translation operable between a pair of Indian languages has been proposed and is tested for translation between Hindi and Malayalam. Experimentations were carried for valid datasets and the training is done with a parallel corpus of five thousand sentences and the translations are adequately intelligible. Thereby the proposed strategy could be validated and would hold good even for a bigger parallel corpus. However local names and domain-specific terms contribute to lower BLEU scores. Furthermore, misspelled words also play a part in the same. As future work, a transliteration module and a domain-specific module could be included in the current system to improve the performances further. The transliteration module will take care of all the local names and terminologies in the language. The proposed model furnishes an average precision of 90.725% with a very low word error rate of 9.125. in the context of Indian Languages which definitely is appreciable. As a further addition, a domain-specific module can also be integrated which will take care of domain-specific terminologies

References:

[1] Das, Aditya Kaustav, Manabhanjan Pradhan, Amiya Kumar Dash, Chittaranjan Pradhan, and Himansu Das. "A Constructive Machine Translation System for English to Odia Translation." In 2018 International Conference on Communication and Signal Processing (ICCSP), pp. 0854-0857. IEEE, 2018.

[2] Sindhu, D.V. and Sagar, B.M., 2016, December. Study on machine translation approaches for Indian languages and their challenges. In 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) (pp. 262-267). IEEE.

[3] Tennage, Pasindu, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. "Neural machine translation for Sinhala and Tamil languages." In 2017 International Conference on Asian Language Processing (IALP), pp. 189-192. IEEE, 2017.

[4] Ahmed, Raihan, Mehedi Al Hasan, and Mohammad Reza Selim. "Aligning Sentences in English-Bengali Corpora." In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), pp. 1-5. IEEE, 2018.

[5] Tholpadi, Goutham, Chiranjib Bhattacharyya, and Shirish Shevade. "Corpus-Based Translation Induction in Indian Languages Using Auxiliary Language Corpora from Wikipedia." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16, no. 3 (2017): 20.

[6] Jadoon, Nadeem Khan, Waqas Anwar, Usama Ijaz Bajwa, and Farooq Ahmad. "Statistical machine translation of Indian languages: a survey." Neural Computing and Applications(2017): 1-13.

[7] Dhariya, Omkar, Shrikant Malviya, and Uma Shanker Tiwary. "A hybrid approach for Hindi-English machine translation." In 2017 International Conference on Information Networking (ICOIN), pp. 389-394. IEEE, 2017.

[8] Kumar, A. and Bansal, N., 2017, September. Machine translation survey for Punjabi and Urdu languages. In 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall) (pp. 1-11). IEEE.

[9] Archana, G. P., V. S. Jithesh, L. B. Remya, and Elizabeth Sherly. "Building a parallel Corpora: Translation issues and remedial case." In 2015 International Conference on Advances in Computing, Communications, and Informatics (ICACCI), pp. 2414-2417. IEEE, 2015.

[10] Deepak, G., Priyadarshini, J.S. and Babu, M.H., 2016, October. A differential semantic algorithm for query relevant web page recommendation. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA)(pp. 44-49). IEEE.

[11] Pushpa, C.N., Deepak, G., Thriveni, J. and Venugopal, K.R., 2015, December. Onto Collab: Strategic review oriented collaborative knowledge modeling using ontologies. In 2015 Seventh

International Conference on Advanced Computing (ICoAC) (pp. 1-7). IEEE.

[12] Deepak, Gerard, Ansaf Ahmed, and B. Skanda. "An intelligent inventive system for personalized webpage recommendation based on ontology semantics." Int. Journal of Intelligent Systems Technologies and Applications 18, no. 1/2 (2019): 115-132.

[13] Deepak, Gerard, and Sheeba Priyadarshini. "A hybrid framework for social tag recommendation using context driven social information." International Journal of Social Computing and Cyber-Physical Systems 1, no. 4 (2016): 312-325.

[14] Deepak, G., Shwetha, B.N., Pushpa, C.N., Thriveni, J. and Venugopal, K.R., 2018. A hybridized semantic trust-based framework for personalized web page recommendation. International Journal of Computers and Applications, pp.1-11.

[15] Pushpa, C.N., Deepak, G., Thriveni, J. and Venugopal, K.R., 2016. A Hybridized Framework for Ontology Modeling incorporating Latent Semantic Analysis and Content based Filtering. International Journal of Computer Applications, 150(11).

[16] Deepak, G. and Priyadarshini, J.S., 2018. A Hybrid Semantic Algorithm for Web Image Retrieval Incorporating Ontology Classification and User-Driven Query Expansion. Advances in Big Data and Cloud Computing, p.41.

[17] Gulzar, Z., Leema, A.A. and Deepak, G., 2018. PCRS: Personalized course recommender system based on hybrid approach. Procedia Computer Science, 125, pp.518-524.

[18] Deepak, Gerard, and S. J. Priyadarshini. "Onto tagger: ontology focused image tagging system incorporating semantic deviation computing and strategic set expansion." Int. J. Comput. Sci. Bus. Inform 16, no. 1 (2016).

[19] CN, Pushpa, Gerard Deepak, Mohammed Zakir, and Venugopal KR. "Enhanced Neighborhood Normalized Pointwise Mutual Information Algorithm for Constraint Aware Data Clustering." ICTACT Journal on Soft Computing 6, no. 4 (2016).

[20] Gerard D and Z. Gulzar. "Ontoepds: Enhanced and personalized differential semantic algorithm incorporating ontology-driven query enrichment." Journal of Advanced Research in Dynamical and Control Systems 9, no. Special (2017): 567-582.

[21] Giri, G.L., Deepak, G., Manjula, S.H. and Venugopal, K.R., 2018. OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 265-275). Springer, Singapore.

[22] Deepak, G . and Priyadarshini, J.S., 2018. Personalized and Enhanced Hybridized Semantic Algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. Computers & Electrical Engineering, 72, pp.14-25.

[23] Santhanavijayan, A., and S. R. Balasundaram. "Multi Swarm Optimization based Automatic Ontology for E-Assessment." Computer Networks (2019)