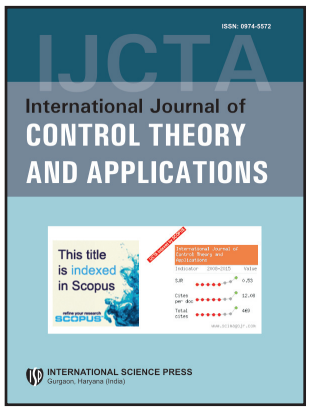

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 43 • 2016

A Model Literature Analysis on Machine Translation System Finding Research Problem in English to Hindi Translation Systems

Priyanka Malviya[a] and Gauri Rao[b]

[a]*Department of Computer Engineering, Bharati Vidyapeeth Deemed University, Collage of Engineering, Pune, India. Email: priyankamalviya1006@gmail.com*
[b]*Prof. Department of Computer Engineering, Bharati Vidyapeeth Deemed University Collage of Engineering, Pune, India. Email: grrao@bvucoep.edu.in*

Abstract: The Statistical Machine Translation (SMT) systems are a computational language and automatically use the software to translate the text and word speeches from one language to another language, have high coverage but very less accuracy. Also, SMT system does not have expressiveness and explain ability. Developing high-quality translation rules is primarily a manual process, requires lots of manual effort and time, and is prone to errors. A machine learning-based framework has the multiple integrated features for the machine translation model pruning. There are many systems existed in statistical machine translation System, where most of the system is not efficient and effective. Researchers have been trying to apply more ideas to evaluate the translation. This paper presents the comprehensive survey of how to evaluate the translation accuracy in statistical machine translation using various methods. There needs to find the translation challenges and find the issues are listed in a relative manner. This paper concludes some future direction that can help the researchers to identify ideas where future works are needed.

Keyword: Accuracy, Machine Learning, Language, Model Pruning, Statistical Machine Translation (SMT), Text translation.

1. INTRODUCTION

The machine translation is a subfield of the computational language and automatically the use of Software to translate the text or speech from source language to target language or also called as machine-aided human translation (MAHT) and which is early interactive for the human. But only Machine Translation alone cannot produce a good and accurate language translation System of Text and Speech because some database also used. To solve this problem with corpus statistical language and neural language techniques is used this combinational gives the leading better translation, handling a difference in linguistic topology, translation of idioms and isolation of anomalies. The machine translation software is effective in domains where formal or formulae language is used. Machine Translation improves the output quality and also achieve better from human interference, the

example-Some system can translate more accurate result by using various text and phrase conversion methods and gives better and proper target language test and word phrases or Speeches.[1]

The human translation process can be explained as follows:

1. Decode the meaning of the source text-First finds the appropriate way to understand the source language text.
2. After that re-coding this meaning in the target text language.

But behind every simple term, there may be very complex operation. The translator should be interpreted and analyze all the features of the text and process which convert the grammar, semantics, syntax and idioms, etc. requires the in-depth knowledge. The programmer's computer challenge is to understand source language text by the machine and to create a new text in the target language. It can reduce the other task and gives the much faster result as compare to other human translation.[1]

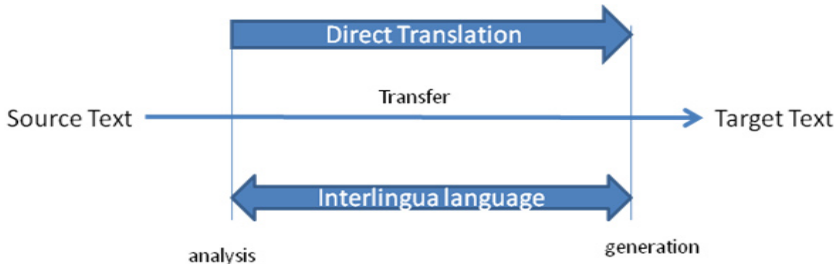

Figure 1: Language Machine translations from source to target[1]

Table 1
Following are the types of machine translation by which we can compare the language translation.[1]

S.N.	*Types of Translation*	*Definition*
1	Rule-based Machine Translation	Rule-based MT is the model based on grammar rules of any language which translate text and speech from the source language to the target language with appropriate standards on both side. The rule based is the dictionary based MT to contain the list of rules in the dictionary format cells rules. It is easy and more efficient than the other translation approaches use for the long term, but it takes more time to complete the task and more databases required to store the results from source to target language. Two terms are important to the rule based translation: (1) Transfer-based MT, it creates translation from an intermediate representation with the simulated meaning of original text sentence and word phrase. (2) Interlingua, translation in one instance.
	Statistical Machine Translation	It is an automatically generate the target text from source language by using various statistical methods based on bilingual text corpora. It provides good result by statistically define the language corpora, but it is still very rare to find the corpora in many languages for the faster translation pairs, which is called the problem challenge in statistical machine translation models. Example –Google translator.
3	Hybrid Machine Translation	Combination of both Rule-based and statistical-based translation gives the approach differ in some ways-rule posts processed by statics and statics guided by rules.
4	Neural Machine Translation	It is the deep-learning approach it is the future for the human language machine translation words in nowadays preference is more given to the process of statistical model approaches.

Challenges for the Machine Translation system[1]:

- It is hard to translate all the words in one language equivalent to the phrase in the target language.
- Word translation problem is because the two languages have the entirely different phrases structures to each other. For example, English language having consonants and vowels only 26 latter combinations and the other Hindi language contains the Devanagari phrases with 52 latter.

- Sometimes lack of one to one association of parts of speech between the source and the target language.
- The way of representation of every language is also different from each other.
- Ambiguity Problem: -The word of the every language may be having more than one meaning (called synonyms) sometimes the group of words and phrases.
- Every time the same rules of grammar are not valid to translate the languages. Translation requires grammar rules and vocabulary knowledge and past used and experiences of the every text or words.
- It is the difficult make the software program to translate the proper text from source to target language. Statistical machine translation is the automatic techniques are to convert the complex programs by understanding the rules of languages which is complex for the human to translate every time.
- Many simulation techniques are present in the today's worlds which automatically translate the language, but it is impossible to achieve the proper and efficient goal and get the accurate results.[1]

INDIA is a multilingual, multicultural country where 22 official languages and approximately 2000 dialects are spoken by different communities[2]. English and Hindi are used for official work in majority of the states of India while state governments predominantly carry out their official work in their regional language such Hindi, Bengali, Hindi, Tamil, Kannada, Telugu, Punjabi, Gujarati, Oriya etc. The people of different states make use of these regional languages for oral as well as written communication. The entire official documents and reports of Union government are published in English or in both English and Hindi. Translating these documents manually into a regional language is very time consuming and costly task. Hence there is need to develop good machine translation (MT) systems in order to establish a Better communication between states and Union governments and exchange of information amongst the people of different states with different regional language English continues to be the link language in India. Machine translation has a much greater significance in breaking the language barrier within the sociological and regional structure[1]. Few MT systems for English to Indian languages for specific domain are developed and the work is still going on[2]. It is a tough task to develop general purpose English to Indian languages Machine Translation systems due to the complex and free-word order nature of different Indian languages. English language has simple, complex and compound sentences. The simple sentences are further sub-divided into Interrogative, Assertive, Negative, and Exclamatory. Developing a tool for each sub-type of simple sentences and integrating them to form a full-fledged MT tool could be a better option. Hindi is a low resource Indian language. The tool for Simple Interrogative English sentences to Hindi has been already attempted[1].We are proposing a system for translating Simple English Assertive sentences into Hindi sentences.

Machine translation has different architectures such as Direct, Transfer Based, Interlingua, Statistical, Example Based and Hybrid. Each of them has its advantages and disadvantages and selection of the approach can be made based on the domain of the application. Proposed research work is an innovation and presents a programmable Machine Translation system.

2. RELATED WORK

Author Present a statistical machine translation model that uses hierarchical phrases that contain sub-phrases Using BLEU as a metric of translation accuracy. It is a modeling challenge by the author to implement hierarchical structure or Syntax-based Machine Translation System because of corresponding syntactic structures with the very disordered parallel corpora. It was difficult to added complexity introduced by hierarchical structures, to reduce

this difficulty author focus on the logical growth of phrase-based approach for improving translation accuracy, which is the core idea of this system. To make this possible various methodology are used in this system.[3]

1. Translation from word to phrase based.
2. Phrased-based is a noisy channel approach.

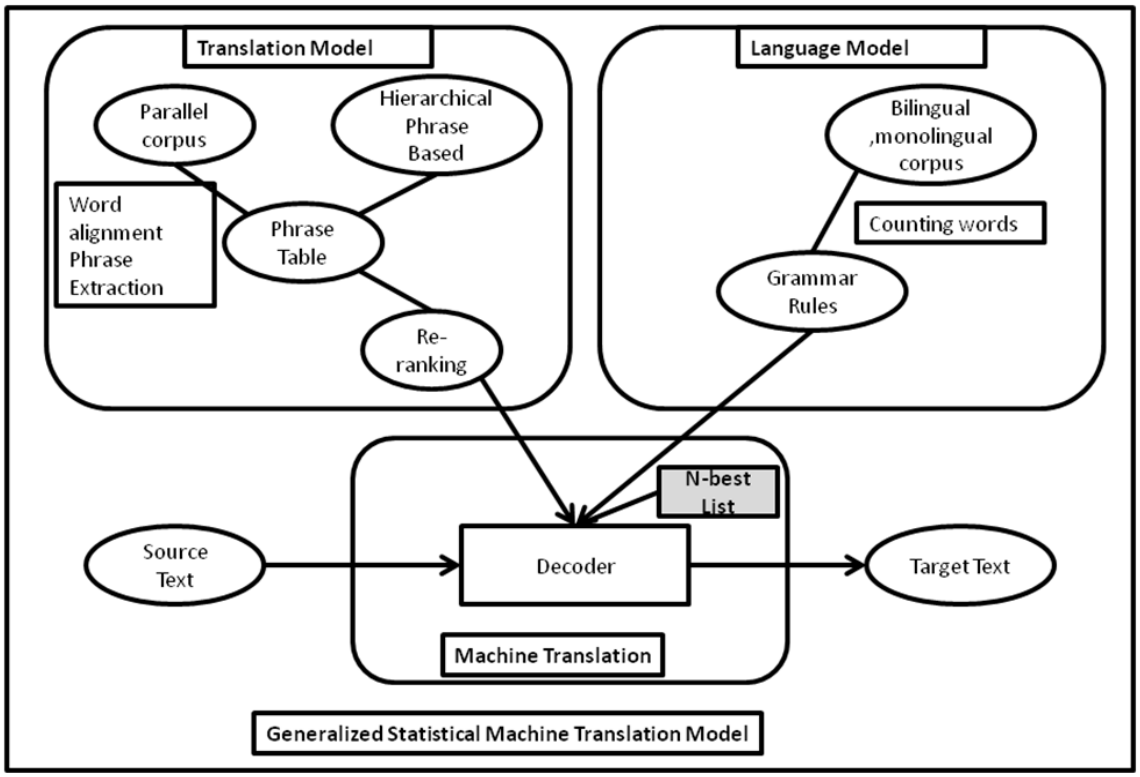

Figure 2: Statistical machine translation (SMT) Model in generalized form translation and language model to convert source test to the target text by using various techniques like phrase based, hierarchical and grammar rules[1], [3]

The system is based on the French to English Translation language and model of this system is based on a Synchronous Context-free grammar, or it is known as a syntax-directed transduction grammar (Leuis and Stearns 1968).The algorithm of the system take 'f' is the French Language's' is Translation and 'D' is the derivations. They describe definition, features, training and decoding by deriving these three terms, and present several phrases as deductive proof System. By using an example of CKY algorithm for CFG's. Model takes 'S' is the start symbol and n is the final length of an input string of 'f' language. The given synchronous CFG convert its French-side grammar into Chomsky normal form and by checking best parse Output using CKY given the final translation in the English Language. The Algorithm gives the O(n3) Complexity same as CKY algorithm[3]. Algorithm:

- Take an array chart taking the character of 3 items in sets.
- Whenever it proves the new thing it adds the item to the appropriate place in chart cell, to reuse the item it stores each item or empties the tuple is not used.
- If true item are added to a cell check the equivalent except and weights, and they will also be merged known as hypothesis recombination.

The hybrid method which they called as cube pruning, it is the compromise between the rescoring and interaction methods. Limitation of this system is the algorithm does not completely search the space of proofs

because it depends on French languages rules. Phrase-based Translation suffers from accuracy, and presented framework is developed for flat structures, not for complex hierarchical structures. The scope of the system is syntactically informed statistical machine translation.[3]

The author introduces a word alignment framework that facilitates incorporation of syntax encoded in bilingual dependency tree pairs. Generate shorter outputs on baseline word alignment. It is work based on the Chinese to the English language. The following are the sub-model describes in this system[4].

1. Two separate processes performed in pipeline
2. Used complex IBM models e.g. IBM model 4 for getting author alignment.
3. Alignment searching processes which assumed that alignment link for each word is made independently.
4. Comparison with traditional explaining word alignment models also shows the disadvantage of this approach.
5. Comparative evaluation is not presented.

This system depends on of tree pairs (PBMT), phrase-based SMT System[3], [4] and the result shows shorter output and improve Chinese to English Machine translation language compare with the baseline word alignment model the result of these model shows up to 10.52% and 5.52% relatively decrease in alignment error rate compared with generate word alignment model and syntax proof descriptive word alignment model. The advantage of this mechanism by which syntactic features may be incorporated contains PBMT system. And not work for larger datasets and more language pairs. The future work of this system adopts the approach for work on large datasets for more languages pairs. Scope it will be better and larger approach and Bootstrap alignment using simple heuristics without relying on complex IBM models.[4]

In the Proposed System author use the Noise as the filtrating principle techniques, and this system shows that when partitioning the bi-phrase tables into several subclasses according to the complexity of the system using Noise to improve the BLEU score which is unreachable using *p*-valve as compare to the similar amount of pruning of the phrase tables. The Algorithm [5] of this system is as follows:

- First, choose the number of threshold levels for the association score β;
- Then build a table of (non-continuous) bi-phrase are based on the bilingual corpus.
- Partition the bi-phrase table U.

Currently, the system is based on the rule-based translation techniques. By using the simulation techniques (statistical machine translation techniques), this system introduced the powerful and flexible way of computing Noise and using the translation relation in the bilingual corpus in the future work.[5]

This system proposed a methodology as a phrase pairs extraction is the core of this system based on the performance of the Phrase-based Machine Translation (PBMT) system, in which Chinese to English language translation used. This system takes the syntactic knowledge for the extraction phrase from the word- based alignment[3], [4], [6] for the PBMT system. As compare to the other system to filter out the entire non-syntactic phrase, this system only filters out non-syntactic phrase first, and the last source word is unaligned. The method of this system is as follows (1) extracting phrase pairs from word-based alignments, it consists of the sequence of source word, and vector of features values represent the contribution for machine translation. PBMT system such as MOSES. (2) Syntactic constraints on phrase pair extraction divide the possible phrases into two types syntactic phrases which contain the word sequence is the more reliable and the non-syntactic phrases which are not reliable. The results of this system yield a 24.38% phrase pair's reduction,

and 0.52 BLEU point improvements compare with another baseline PBMT system with the full-size of phrase tables.[6]

In this system, the author introduced a novel filtration term to restrict the rule extraction for the hierarchical phrase-based translation model[3]. Where a BLEU but relaxed well-formed dependency can filter out the bad rules. The other feature also added in this system which shows the regularity for source/target word dependency. The lexical weighting concept is used to check the proper words of the phrase to translate each other.

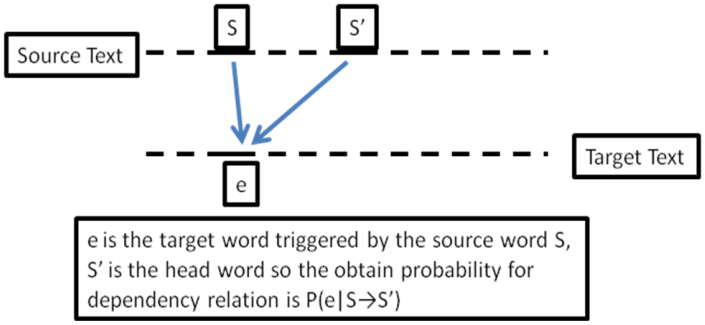

Figure 3: Relaxed well formed dependency structure[7]

The Algorithm of this system is as follows: Take the source words align with target name e, according to IBM model lexical translation probability is P (e/s). The result of this system shows 40% of rules extraction with translation performance improvements, and another feature brings to improve the baseline system. This system aim is to reduce the hierarchical table by not taking an advantage of any linguistic information's.[7]

In this System, the author presented a method for creating trees using higher order-lexical dependencies by using phrase re-ranking framework, this approach models on re-rank N-best lists and forests-Based on dependency features. This system evaluates the mechanism by which syntactic features may be integrated. The methods [8]of this System are-

1. The higher-order lexical dependencies can optimize greater improvements in continues parsing performance then commonly used first-order lexical dependencies.
2. This model increases the dependency accuracy of all the previous trees results as well.
3. All the results combined us on the Chinese Treebank; this method is language independent.

It can be modified to languages with present continues order of trees with labeled dependency tree. Which can provide searching using N-best list or parse forests, the result of this method provides as highest F1 score reaches 85.74% to get more accuracy to all previously reported state-of-art systems. Dependency Parses trees having three type[8].

- Dependency-driven constitute parsing which is not affected on the F1 score.
- Dependency-constrained constitutes parsing-the constitute tree may be pruned directly.
- Dependency-based constitute parsing is first order lexical dependency only.

The first order lexical dependency is quite limited and it may lose much information about the grammatical relation between words performance is also limited as well. To overcome this drawback, this system evaluates continuous tree using higher-order lexical dependencies within a parse re-ranking framework.[8]

The author was classifying bilingual translation pair extracted from aligned corpora. The classification process is the primary core method for this System for this Support Vector Machine (SVM) is used. The

motivation of this system was taken by the fact that automatically extracted the translation as compare or equal after human validation are used for iteratively aligning, extracting and validating are used for translation pairs. This can be possible by auto filtration of appropriate and inappropriate translation pairs to human translation. An example of this method takes input as 1,000 entries validated per hours per validation, by saving the time and progressively improving alignment and extraction quality and to improve the translation quality. Results had shown the larger training set the quality of checked pairs of translation. In this system language taken as English to Portuguese machine translation language, the accuracy calculated for automatically extracted bilingual pairs is seen to be over 85%.And this method supports the input of lexicon translation value using SVM as a source value and target value as co-occurring frequencies as a measure to validate the bilingual pairs. This experiment using lexicon extraction and validation using Moses (Moses is the tool) by using the translation for larger parallel corpora and validation bilingual dictionary needs to be done. The future work of this system is machine learning to classifying translation equivalent to English to the Hindi language for these suffix based features will be examined.[9]

In this existing system, the author evaluate two methods for translation model pruning[10]

- One for the significance test
- One based on relative entropy.

The main core of this system is to improve relative entropy pruning using statistical machine translation Approach, and effective at removing phrase pairs that are results of misalignments, redundancy and found other phrase pairs using proposed methods. The algorithm of this system take first relative entropy pruning for translation models then significance pruning of phrase tables [3], [4] based on statistical machine system can translate the phrases pairs from source to target. The result of this system gives the accuracy improving 20 % relative entropy by calculating the phrases and words pairs by MOSES.[10]

This system work on the French to English, Dutch to English, French to the Russian language based on the statistical Machine Translation, the proposed system have the phrase table pruning method addressed with the ad-hoc way using heuristics approaches. The old methodology idea taken by this system contains Absolute pruning-which take single phrase pairs independent on the phrase in the phrase tables. Count-based pruning and the probability-based pruning are the two types of absolute pruning which function on count and probability using the threshold value. Relative pruning-prune all the occurrence of the source phrase. Threshold pruning and the histogram pruning are the two approaches of relative pruning, threshold pruning discards the phrase that is worse than the best for target phrase from the given source phrase, and the histogram pruning is an alternative to old threshold pruning approach. The above work is not doing well because the phrase table size for the pruning is bigger than other and should be important. So, the system proposed novel-entropy based criterion and with phrase table pruning[10], [11] and also performed systematic tentative comparisons of the existing system methods. The experiments of this system carried out four language pairs under small medium and large data conditions, so the conclusion of this system contain, Probability-based pruning, Count-based pruning as well as significance-based pruning, Entropy-based Pruning gives large saving in phrase table size, Novel entropy-based pruning is also achieved the same BLEU score with an only half in the numbers of phrase.[11]

3. PROBLEM DEFINITION

To enhance the process of machine translation proposed system introduced graph based pruning approach. Achieve optimized result and accurate translation for English to Hindi Translation.

4. CONCLUSION

Above research, the article is literature review on Programmable Machine translation paradigm designed, which present the open challenges and problems to solve the Machine translation for the source to target language to evolving structured and growing with the human translator for efficient and effective statistical machine translation. The Large scope of work is present in these approach section I shows the introduction and challenges of Machine Translation and section II show the related literature survey to solve the problems in future.

Acknowledgment

To prepare the Literature Survey Paper of-" A Model Literature Analysis on Machine Translation System Finding Research problem in English to Hindi Translation systems" has been prepared by Miss. Priyanka Malviya and Prof. Gauri Rao.

I would like to thank my faculty as well as my whole department, parents, and friends for their support and confidence in me. I have obtained a lot of knowledge during the preparation of this document.

REFERENCES

[1] N. Language, P. Home, A. Contact, M. Site, M. A. P. Search, M. Anti, R. P. Held, B. Threatens, L. Computers, S. Your, C. Abbreviations, B. Bots, G. Grammar, L. Branches, I. Machine, T. Modality, G. Nlp, O. Parsing, P. O. S. Tagging, P. References, and S. Tamil, "Natural Language Processing," pp. 2–5, 2017.

[2] P. Dungarwal, R. Chatterjee, A. Mishra, and A. Kunchukuttan, "The IIT Bombay Hindi ⇔ English Translation System at WMT 2014," 2014.

[3] D. Chiang, "Hierarchical phrase-based translation," Comput. Linguist. Vol. 33, pp. 201–228, 2007.

[4] Y. Ma, S. Ozdowska, Y. Sun, and A. Way, "Improving word alignment using syntactic dependencies," in Proc. 2nd ACLWorkshop Syntax andStructure in Statist. Translat., Columbus, OH, USA, Jun. 2008, pp. 69–77.

[5] N. Tomeh, N. Cancedda, and M. Dymetman, "Complexity-based phrase-table filtering for statistical machine translation," in Proc. MT Summit XII, Ottawa, ON, Canada, Aug. 2009.

[6] 2011H. Cao, A. Finch, and E. Sumita, "Syntactic Constraints on Phrase Extraction for Phrase-Based Machine Translation," in Proc. SSST-4, 4th Workshop Syntax and Structure in Statist. Translat., Beijing, China, Aug. 2010.

[7] Z. Wang, Y. Lü, Q. Liu, and Y. S. Hwang, "Better filtration and augmentation for hierarchical phrase-based translation rules," in *Proc. ACL Conf. Short Papers*, Uppsala, Sweden, Jul. 11–16, 2010, pp. 142–146.

[8] Z. W. a. C. Zong, "Parse Reranking based on higher-order lexical dependencies," in Proc. 5th Int. Joint Conf. Nat. Lang. Process., Chiang Mai, Thailand, Nov. 8–13, 2011, pp. 1251–1259.

[9] K. Kavitha, L. Gomes, and G. P. Lopes, "Using SVMs for Filtering Translation Tables for Parallel Corpora Alignment," in Proc. EPIA.

[10] L. Wang, T. Nadi, X. Guang, B. Alan, and T. Isabel, "Improving relativeentropy pruning using statistical significance," in Proc. 25th Int. Conf. Comput. Linguist. (Posters), Mumbai, India, Dec. 2012, pp. 713–722.

[11] R. Zens, D. Stanton, and P. Xu., "A systematic comparison of phrase table pruning techniques," in Proc. Joint Conf. Empir. Meth. Nat. Lang. Process. Comput.Nat. Lang. Learn., Jeju Island, Korea, Jul. 2012, pp. 972–983.

[12] Andrei Alexandrescu, Katrin Kirchhoff, "Graph based Learning for Statistical Machine Translation," in Human Language technologies: The 2009 Annual Confrence of the North American Chapter of the ACL, pages 119-127 Boulder, Colorado, June 2009.

[13] S. B. Kulkarni and K. V Kale, "Linguistic Divergence Patterns in English to Marathi Translation," *Int. J. Comput. Appl.*, Vol. 87, No. 4, pp. 21–26, 2014.

[14] R. Technol, "Divergence patterns between English and Sanskrit machine translation," No. August, 2016.

[15] R. Sinha and A. Thakur, "Divergence patterns in machine translation between Hindi and English," ... *Mach. Transl. summit (MT Summit X)*... pp. 346–353, 2005.

[16] D. Bhalla, M. Tech, N. Joshi, D. Ph, I. Mathur, and M. Sc, "Divergence Issues in English-Punjabi Machine Translation. pdf," Vol. 14, No. October, pp. 73–83, 2014.

[17] L. T. Centre, "DIVERGENCE PATTERNS IN HINDI - MALAYALAM MACHINE TRANSLATION," Vol. 15, No. December, pp. 12–17, 2014.

[18] B. S. Gillon, "Sanskrit Computational Linguistics," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Vol. 5406, No. November, pp. 98–105, 2009.

[19] M. Computing, "Inflection Rules for English to Marathi Translation," Vol. 2, No. April, pp. 7–18, 2013.

[20] R. M. K. Sinha, "Divergence patterns for Urdu to English and English to Urdu Translation.," No. 1, pp. 21–28, 2011.

[21] G. Rao, C. Agarwal, S. Chaudhry, N. Kulkarni, and D. S. H. Patil, "Natural language query processing using semantic grammar," *Int. J. Comput. Sci. Eng.*, Vol. 2, No. 2, pp. 219–223, 2010.

[22] G. Rao and R. Based, "THREE DIMENSIONAL VIRTUAL ENVIRONMENT FOR Address for Correspondence," No. Ii, 2011.

[23] R. S. Anami, "Automated Profile Extraction," No. x, pp. 208–211, 2014.

[24] R. S. Anami and G. R. Rao, "Automated Profile Extraction and Classification with Stanford Algorithm," No. 7, pp. 67–71, 2014.

[25] P. Dungarwal, R. Chatterjee, A. Mishra, and A. Kunchukuttan, "The IIT Bombay Hindi ⇔ English Translation System at WMT 2014," 2014.

[26] M. Tech, "Lexicon-Based Approach and Addressing," Vol. 90, No. 1, 2016.

[27] S. Patil and P. G. Rao, "Web Page Template Generation and Detection of Non- Informative Blocks Using Trinity," Vol. 5, No. 3, pp. 70–72, 2016.