Survey of Machine Translation for Indian Languages to English and Its Approaches

Namrata G Kharate[1] , Dr. Varsha H. Patil[2]

[1]Department of Computer Engineering, VIIT,Pune, Maharashtra, India

[2]Head of Department, Department of Computer Engineering, MCOERC, Nashik, Maharashtra, India

ABSTRACT

This paper presents a survey of Machine translation system for Indian Regional languages to English language. Machine translation is important for breaking language barrier. Out of 1.3 billion India population, hardly 10% people can speak, read and write English with varied degree of proficiency. With the advent of Information Technology many documents and web pages are coming up in a local language so there is a large need of good MT systems to address all these issues in order to establish a proper communication between states and union governments to exchange information amongst the people of different states. This paper discusses the various approaches which have been applied in translation systems for Indian languages to English languages. Some of the important Indian language translation systems implemented with these techniques along with their capabilities and limitations are also discussed.

Keywords: Machine translation, Indian Languages, Rule-based, Statistical, Knowledge-based, Hybrid Machine Translation, Interlingua-based MT

I. INTRODUCTION

Machine translation is one of the central areas of Natural language processing (NLP). Machine translation (MT) is important for breaking the language barrier among the people and for inter-lingual communication where translation is done from source language to target language. Many researchers, Institutions and organizations in India have started working on MT systems for Indian languages and have gained satisfactory results [3] Communication has been a vital part of the life of humans from the beginning of time. With about 71 million Marathi speaking people and varied works in Marathi literature and novels calls for translation [1]. Languages are the tools for effective communication. Marathi is one of the 22 official languages of India. Many official documents and lot of information these days are available usually in the English language. Existing documents that are currently in the Marathi language need to be translated to English for their

widespread use. Manual translation is very costly and time consuming and hence there is a need to have an automated translation system which would do the language translation in an effective way. Also, there is not much work done so far for Marathi to English language and for Indian languages to English language also. The major challenge is the structural difference between Indian languages and English language. English follows are a Subject-Verb-Object grammar structure while Marathi language follows Subject-Object-Verb grammar structure, relatively of free word order and has large number of inflections. Hence its translation to English or vice a versa is a challenging task.

II. MACHINE TRANSLATION APPROCHES

Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as

English) to another (such as Spanish). It is translation of text by a computer, with no human involvement. Many machine translation systems across the world have already been developed for mostly used natural languages. Generally MT is classified into various approaches. Figure 1 depicts the various machine translation approaches.
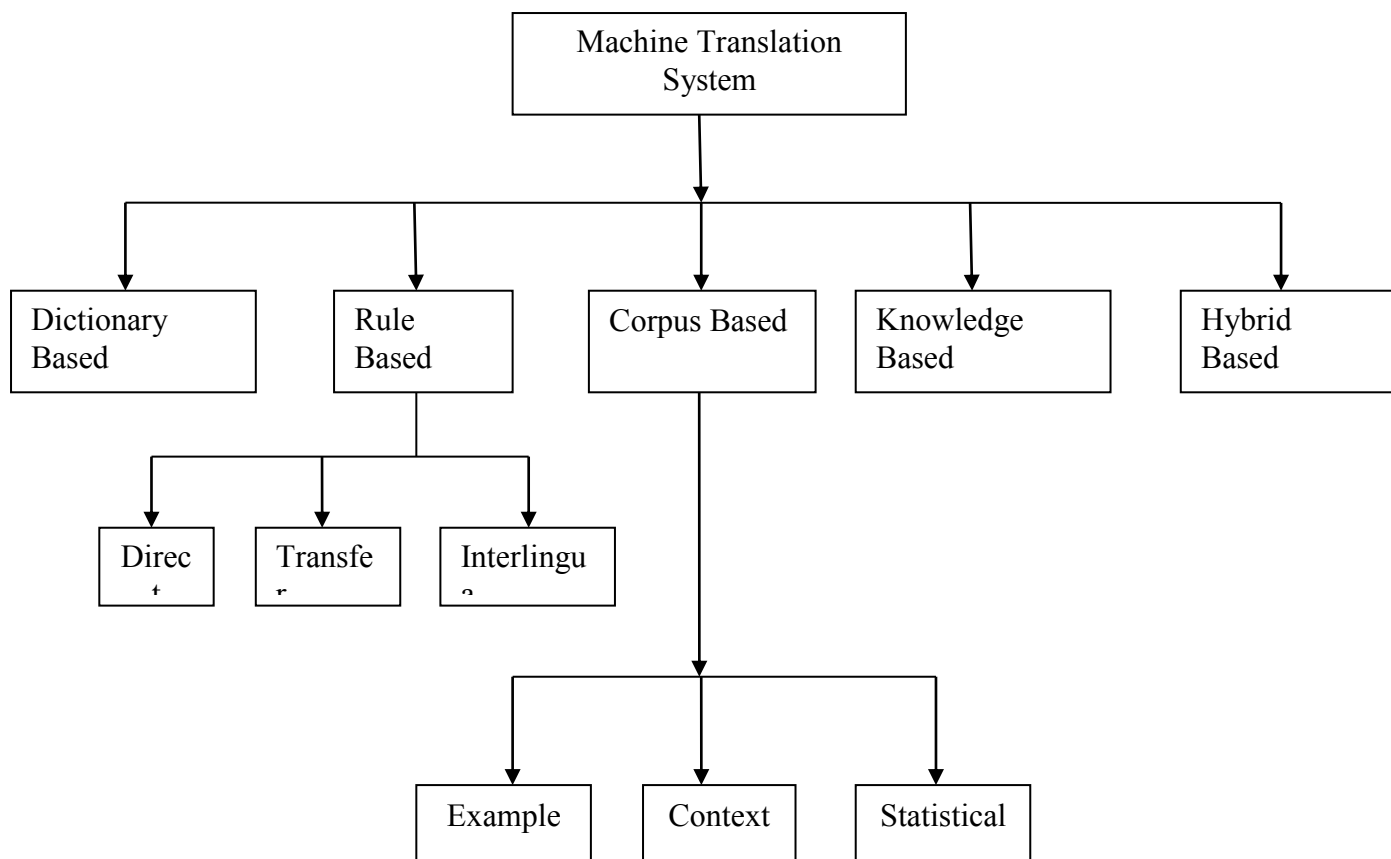
Figure 1. Machine Translation Approaches

1. Dictionary based Translation

This Machine Translation method is based on dictionary entries, which means that the translation is done word to word, usually without much correlation of meaning between them. Dictionary lookups may be done with or without morphological analysis or lemmatization. This approach to Machine Translation is probably the least sophisticated but it is ideally suitable for the translation of long lists of phrases on the sub-sentential (not a full sentence) level. Examples are Inventories or simple catalogs of products and services. It can also be used to expedite manual translation, if the person carrying it out is fluent in both languages and therefore capable of correcting syntax and grammar

2. Rule-based Machine Translation

In this approach linguistic rules are build according to source language and target language. RBMT systems are based on adjusting the structure of the given input sentence with the structure of the target output sentence, preserving their unique meaning. It parses the source language sentence and produces an intermediate representation such as parse tree or some abstract representation. It uses large collection of manually developed rules used for mapping source language into target language text. These rules can be modified to improve translations. Based on the

intermediate representation, these systems are further categorized as Direct, Transfer and Interlingua. Figure 2 depicts the translation schemes known as the Vauquois triangle.
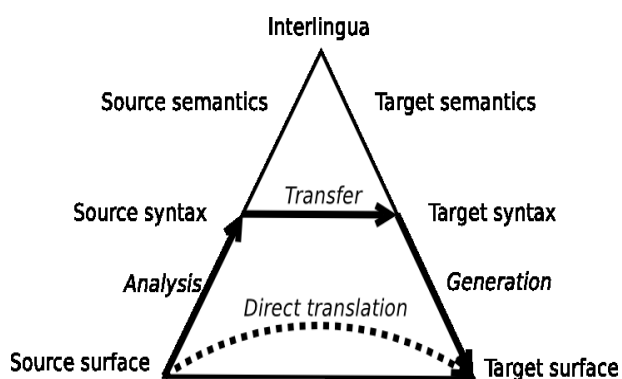

Figure 2. The Vauquois triangle

2.1 Direct Translation

Direct Machine Translation is the one of the simplest machine translation approach in which a direct word to word translation is done with the help of a bilingual dictionary without considering the semantics of source and target language.

2.2 Transfer Based Translation

A Rule-Based Machine Translation (RBMT) system consists of collection of various rules, called grammar rules, a bilingual lexicon or dictionary, and software programs to process the rules.

2.3 Interlingua Based Translation

In this approach, the translation consists of two stages, where the source Language (SL) is first converted in to the Interlingua (IL) form. The main advantage of Interlingua approach is that the analyzer and parser of SL is independent of the generator for the Target Language (TL).

3. Corpus-based Machine Translation

3.1 Statistical Approach

Statistical Machine Translation (SMT) is a data-oriented statistical framework which is based on the knowledge and statistical models which are extracted from bilingual corpora. In this MT, bilingual or multilingual corpora of the languages are required. In SMT, a document is translated according to the probability distribution function which is indicated by p(e|f). Finding the best translation is done by picking the highest probability, as shown in Equation 1.

$$e= \text{argmax } p\,(e \mid f) = \text{argmax } p\,(f \mid e)\,p(e) \ldots \ldots \ldots \ldots (1)$$

3.2 Example-based translation

Basic idea of this MT is to reuse the examples of already existing translations. An example-based translation uses a bilingual corpus as its main knowledge base and it is essentially translation by analogy.

3.3 Context-based Machine Translation (CBMT)

It is being developed as a corpus-based method that requires neither rules nor parallel corpora. The CBMT requires an extensive monolingual target text corpus, a full-form bilingual dictionary, and a smaller monolingual source text corpus some times.

4. Knowledge-Based MT

Knowledge-Based Machine Translation (KBMT) requires complete understanding of the source text prior to the translation into the target text. KBMT is implemented on the Interlingua International Journal on Natural Language

5. Hybrid Machine Translation

Machine Translation architecture hybridizes a rule-based system, with regular phrase-based Statistical Machine Translation. The hybrid translation process is guided by the rule-based engine and a set of partial candidate translations provided by SMT subsystems. SMT subsystem is used to enrich the tree-based representation. The final hybrid translation is created by choosing the most probable combination among the available fragments with a statistical decoder in a monotonic way. The hybrid model is applied to a pair of distant languages Spanish and Basque and evaluation (both automatic and manual) shows that it

significantly outperforms the best SMT systems on out of domain data [22][23]

III. LITERATURE SURVEY

Although the significant work is not carried out for Marathi to English language translation or Indian languages to English language translation but many researchers and organizations have started the research in this area and started building machine translation tools/systems in recent past in India. In this section we look at some Indian Languages to English language machine translation project.

3.1 Translation system for Hindi to English translation

3.1.1 Hindi-English transfer based MT system (2015)
Shashi Pal Singh, Ajai Kumar, Akanksha Gehlot and Vaishali Sharma have developed a system which is based on transfer based machine translation approach. This system comprised of modules such as pre-processing, tokenizer, morphological analysis, parsing and post-processing. For parsing CYK (cocke younger kasami) algorithm is used. It follows bottom up approach. It works on Context free grammar known as parallel dynamic syntax analysis algorithm. System does not handle the sentence consisting of idioms, complex, compound and other form of sentences in perfect and perfect continues tenses [4].

3.1.2 Hindi-English MT system using Effective selection in multi model SMT (2014)
In this, stastical machine translation system using multiple translation models is used. This system uses multiple strategies for translation and chooses one translation as final output. This selection is done using prediction stratergies. System uses ILCI corpora which contains parallel sentences for 11 languages from health domain with hindi as their source language.25000 parallel sentences are used for experiments [5]. This system trained two models phrase model and hierarchical model.

3.1.3 IIIT Bombay Hindi <--> English MT system at WMT (2014)
In WMT 2014, statistical machine translation approach is used. English to Hindi and Hindi to English language translation is carried out. Phrase based approach used for Hind to English translation and factored based approach for English to Hindi. Pre-ordering Hindi source corpus is done by Shallow parser. Supertag factors are used for better translation of complex sentence. Accurately noun inflections are generated in Hindi using number case factors [6].

3.1.4 Hinglish MT System (2004)
Sinha and Thakur developed Hinglish - a machine translation system for pure Hindi to pure English forms. It incorporates additional level to the existing English to Hindi translation (AnglaBharti-II) and Hindi to English translation (AnuBharti-II) systems developed by Sinha. The system has produced satisfactory acceptable results in more than 90% of the cases. The system is not capable of resolving the meaning of polysemous verbs due to a very shallow grammatical analysis used in the process [7].

3.1.5 GB Theory Based Hindi to English Translation System (2009)
Example Based System makes use of Government and Binding (GB) theory for undertaking translation. It consists of a parsing module and generating module.

3.1.6 ANUBHARTI-I (2003)
Anubharti (Sinha, 2004) is developed using a hybridized example-based machine translation approach i.e. a combination of example-based, corpus-based approaches and some elementary grammatical analysis. The example-based approaches follow human-learning process for storing knowledge from past experiences and to be used it in future. In Anubharti, the traditional EBMT (Gupta and Chatterjee, 2003) approach has been modified to reduce the requirement of a large example-base. The modification in traditional EBMT is achieved by generalizing the constituents and replacing them with abstracted form from the raw examples. The

abstraction is achieved by identifying the syntactic groups. Matching of the input sentence with abstracted examples is done based on the syntactic category and semantic tags of the source language structure. The architectures of both AnglaBharti and AnuBharti, have undergone a considerable change from their initial conceptualization. In 2004 these systems were named as AnglaBharti-II and AnuBharti-II respectively. AnglaBharti-II uses a generalized example-base for hybridization besides a raw example-base and the AnuBharti-II makes use of Hindi as source language for translation to any other language. The generalization of the example-base is dependent upon the target language [21][9].

3.1.7 ANUBHARTI-II (2004)

R M K Sinha developed a MT system using Generalized Example-Base (GEB) along with Raw Example-Base (REB) MT approach for hybridization. The combination of example-based approach and traditional rule-based approach is used in this system. The example based approach emulates human-learning process for extracting knowledge from past experiences and to be used in future. The source language is Hindi. The inputted Hindi sentence is converted into a standard form to handle the word-order variations. The Hindi sentences converted into standard form are matched with a top level standard form of example-base. If no match is found then a shallow chunker is used to fragment the input sentence into small units and then they are matched with a hierarchical example-base. The small chunks obtained by shallow chunker are translated and positioned by matching with sentence level example base [21][9].

3.2 Translation system for Sanskrit to English translation
3.2.1 TranSish (2014)

Pankaj Upadhyay, Umesh Chandra Jaiswal, Kumar Ashish developed a system using Rule based approach. With the help of Artificial Intelligence system provide an interface which converts Sanskrit sentences to English .In this machine translation

Sanskrit sentence as a source language is input through an interface called transish interface.Parsing is done with the help of Unicode to corresponding English characters.Transish algorithm used for post-processing.there are six tenses in sansnkrit,out of that Lat lakara (present tense)is focused in system.

3.3 Translation system for Punjabi to English translation
3.3.1 Statistical MT Based Punjabi to English Transliteration System for Proper Nouns (2013)

The proposed system is divided to two parts – learning and transliteration. System is based on Statistical machine translation approach. The system Accuracy depends on data stored into the database and overall accuracy of the system is measured up to 97%.N-Gram approach is used to transliterate the proper nouns of Gurumukhi script into English.

3.3.1 Rule Based Machine Translation of Noun Phrases (2010)

Kamaljeet Kaur Batra and G S Lehal developed system based on transfer based approach. The system has analysis, translation and synthesis component. A Punjabi morph analyzer is being used for analyzing the exact grammatical structure of the word. Overall accuracy of translation is 85.33%.

3.4 Translation system for urdu to English translation
3.4.1 Urdu to English Machine Translation using Bilingual Evaluation Understudy (2013)

This system is based on Corpus machine translation approach and makes use of three main techniques of machine translation: Rule based Machine translation, Statistical Machine translation and Example Machine translation. Methodology of each system is explained and found their comparison based on their respective outputs using BLEU. The EBMT approach produced accuracy of 84.21% whereas the accuracy of the online SMT system is 62.68%.

3.5 Translation system for Malayalam to English translation

3.5.1 Design & Development of a Malayalam to English Translator (2012)

Latha R Nair, David Peter & Renjith P Ravindran have developed system based transfer based approach.The system consists of a pre-processor for splitting the compound words, a morphological parser for context disambiguation and a bilingual dictionary. The system is designed using artificial intelligence techniques and can easily be modified to build translation systems for other language pairs

3.5.2 Malayalam to English Machine Translation (2014)

System is based on General Example machine translation approach. Consist of 3 phases: Acquisition, Matching & Recombination. Each Malayalam fragment is searched into corpus.About 75% of the test gives good quality translation. The system works well for the all simple sentences in their 9 tense forms, their negatives and question form.

3.6 Translation system for Manipuri to English translation

3.6.1 Bidirectional MT Systems using Morphology and Dependency Relations (2010)

Thoudam Doren Singh ,Sivaji Bandyopadhyay have developed a system based on Statistical based approach. The systems are evaluated using automatic scoring techniques BLEU and NIST. Uses morphology and dependency relations and domain specific parallel corpus of 10350 sentences from news for training purpose. The System is tested with 500 sentences.

3.7 Translation system for Marathi to English translation

3.7.1 Marathi to English Sentence Translator for Simple Assertive and Interrogative Sentences (2016)

System is based on Rule-based machine translation approach. This system comprised of modules such as source language parsing, Bilingual Lexicon and target language generator. Uses shallow parser for parsing of Marathi language. Rearrangement algorithms used for target language generator. Disambiguation of nouns and verbs is provided. It can be extended for complex and compound sentence

3.8 Google translator (2007)

Google Translate is a free multilingual machine translation system. It is a multilingual service supports 91 languages. The system depends on a solid corpus. It offers a web interface, mobile interface for android and iOS. The accuracy of translation is better to understand the translated text. It uses Statistical Machine Translation i.e. machine translation in which translation is generated using statistical translation models and parameters of which are derived from the analysis of bilingual text corpora. If corresponding word is not found in the text corpora, accurate translation is not obtained. Moreover the Google translate does not check the syntax of the given sentence.

The following table describes some of the Machine Translation systems developed for Indian languages to English language translation.

Table 1. Machine Translation Systems for Indian Languages to English

SN	Machine translation system	Year	Languages pair	MT Approach used	Observations
1	Hindi to English transfer based Machine Translation System[4]	2015	Hindi -English	Transfer Based	-CYK algorithm used for parsing -No translation for sentences consisting idioms, complex, compound sentences
2	Hindi to English Machine Translation: Using Effective Selection in Multi-Model SMT[5]		Hindi -English	Statistical	-Trained two models, phrase model and hierarchical models -Used 25000 parallel sentences -Need to integrate few more linguistic and statistical features
3	The IIT Bombay Hindi –English Translation system[6]	2014	Hindi -English	Statistical Phrase based and factored based	-Uses preprocessing and post processing components. -Use of supertag factors for better translation of structurally complex sentence -Use of shallow parsing for preordering Hindi source
4	Hinglish MT System[7]	2004	Hindi -English	Example based	-Based on Anubharti-II & Anglabharti-II. -It produces satisfactory results in more than 90 cases. -It performs shallow grammatical analysis
5	GB Theory Based Hindi To English Translation System[8]	2009	Hindi -English	Example Based	-System makes use of Government and Binding (GB) theory for undertaking translation. -It consists of a parsing module and generating module

6	ANUBHARTI-I[9]	2003	Hindi -English	Hybrid	-Combination of example based, corpus based & some grammatical analysis -It reduces the requirement of large example base and it depends on target language
7	ANUBHARTI-II [9]	2004	Hindi -English	Hybrid	-It emulates human-learning process for storing knowledge from past experience to use it in future. -Shallow chunker is used for fragmentation of input sentences.
8	TranSish [10]	2014	Sanskrit-English	Rule-based	-The system provide an interface which converts Sanskrit sentences to English with a rule based model of parser and a semantic Mapper.
9	Statistical Machine Translation Based Punjabi to English Transliteration systems for proper noun [11]	2013	Punjabi-English	Statistical	-The system is divided to two parts – learning and transliteration. -System accuracy depends on data stored in the database and overall accuracy of the system is measured up to 97%.
10	Rule Based Machine Translation of Noun Phrases from Punjabi to English[12]	2010	Punjabi-English	Transfer-based	-The system has analysis, translation and synthesis component. -A Punjabi morph analyzer is being used for analyzing the exact grammatical structure of the word. -Overall accuracy of translation is 85.33%.

11	Urdu to English Machine Translation using Bilingual Evaluation Understudy [13]	2013	Urdu-English	Corpus- based	-Explained methodology of each system and their comparison based on their respective outputs using BLEU. -The EBMT approach produced accuracy of 84.21% whereas the accuracy of the online SMT system is 62.68%.
12	Malayalam To English Translator[14]	2012	Malayalam-English	Transfer-approach	-The system consists of a pre-processor for splitting the compound words, a morphological parser for context disambiguation and a bilingual dictionary. -The system is designed using artificial intelligence techniques and can easily be modified to build translation systems for other language pairs.
13	Malayalam to English Machine Translation [15]	2014	Malayalam-English	Example based	-Consist of 3 phases: Acquisition, Matching & Recombination. -Each Malayalam fragment is searched in to corpus. -About 75% of the test gives good quality translation.
14	Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations[16]	2010	Manipuri-English	Statistical approach	-Uses morphology and dependency relations -The system uses domain specific parallel corpus of 10350 sentences from news for training purpose -System is tested with 500 sentences.
15	Marathi to English sentences translator for simple Assertive and interrogative sentences[2]	2016	Marathi-English	Rule Based	-Uses shallow parser for parsing -Disambiguation of nouns and verbs is provided -It can be extended for complex and compound sentence

16	Google Translator[17]	2007	IL-IL	Statistical based	-Google Translate is a multilingual service supports 91 languages. -The system depends on a solid corpus. -The accuracy of translation is better to understand the translated text

IV. CONCLUSION

The present paper discusses the various language translation systems developed in the India which follows different approaches. The systems are developed for various Indian languages to English language which include various language pairs such as Hindi-English, Sanskrit-English, Malayalam-English, Manipuri-English, Punjabi-English, Urdu-English and Marathi-English.

Much of the work is required for translating Marathi sentences to English sentence. The prominent work in the area of Marathi to English translation is restricted to assertive and interrogative sentences. More over the work is based on shallow parsing that needs to be enhanced.

V. REFERENCES

[1]. http://www.censusindia.gov.in/2011-common/censusdataonline.html

[2]. G V Garje, Akshay Bansode, Suyog Gandhi and Adita Kulkarni. Article: Marathi to English Sentence Translator for Simple Assertive and Interrogative Sentences. International Journal of Computer Applications 138(5):42-45, March 2016.

[3]. Amruta Godase, Sharvari Govilkar, "MACHINE TRANSLATION DEVELOPMENT FOR INDIAN LANGUAGES AND ITS APPROACHES" International Journal on Natural Language Computing (IJNLC) Vol. 4, No.2,April 2015.

[4]. Shashi Pal Singh , Ajai Kumar, Aakansha Gehlot and Vaishali Sharma," Hindi to English Transfer based Machine Translation Syetm". International Journal of Advanced Computer Reaserch, ISSN(print):2249-7277 ISSN (online):2277-7970 Volume-5 Issue-19 June 2015.

[5]. Kunal Sachdev, Hrishabh Srivastava, Sambhav Jain,Dipti Mishra Sharma "Hindi to English Machine Translation:Using Effective Selection in Multi-model SMT " LREC 2014.

[6]. Piyush Dungarwal,Rajen Chattarjee,Abhijit Mishra,Anoop Kunchunkuttan,Ritesh shah,Pushpak Bhattacharyya "The IIT Bombay Hindi <- -> English Translation System at WMT 2014" ,Procedding of 9th workshop on Statistical machine translation,Pages90-96,Baltimore,Maryland USA .June26-27,2014

[7]. http://www.academia.edu/7986160/Machine_Translation_of_Bilingual_Hindi-English_Hinglish_Text

[8]. Choudhary,A. Singh, M. (2009) "GB theory based Hindi to english translation system", Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference PP.293 – 297

[9]. Sudip Naskar & Shivaji Bandyopadhyay, (2005) "Use of Machine Translation in India: Current status" AAMT Journal, pp. 25-31

[10]. Pankaj Upadhyay, Umesh Chandra Jaiswal, Kumar Ashish, " TranSish: Translator from Sanskrit to English-A Rule based Machine Translation" , International Journal of Current Engineering and Technology E-ISSN 2277 –

4106, P-ISSN 2347 - 5161 ©2014 INPRESSCO®, All Rights Reserved Available at http://inpressco.com/category/ijcet

[11]. Pankaj Kumar and Er.Vinod Kumar, 2013, "Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns", In International Journal of Application or- Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 8, August 2013 ISSN 2319 – 4847, pp . 318-320

[12]. Kamaljeet Kaur Batra and G S Lehal, 2010, "Rule Based Machine Translation of Noun Phrases from Punjabi to English", In IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010 ISSN (Online): 1694-0814, pp. 409-413

[13]. Asad Habib, Asad Abdul Malik ,Kohat University of Science and Technology, Kohat, Pakistan , 2013, "Urdu to English Machine Translation using Bilingual Evaluation Understudy" International Journal of Computer Applications (0975 – 8887) Volume 82 – No 7, November 2013, pp. 5-12

[14]. Latha R Nair, David Peter & Renjith P Ravindran, 2012, "Design and Development of a Malayalam to English Translator-A Transfer Based Approach", International Journal of Computational Linguistics (IJCL), Volume (3) : Issue (1) : 2012

[15]. Anju E S, Manoj Kumar K V, 2014, " Malayalam To English Machine Translation: An EBMT System", IOSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719 Vol. 04, Issue 01 (January. 2014), ||V1|| PP 18-23

[16]. Thoudam Doren Singh,Sivaji Bandyopadhyay "Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations", Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 83–91,COLING 2010, Beijing, August 2010.

[17]. https://translate.google.co.in

[18]. G.V.Gajre, G.Kharate, H. Kulkarni (2014), "Transmuter: An approach to Rule-based English to Marathi Machine Translation" , International Journal of Computer Applications (0975 – 8887) Volume 98 – No.21, July 2014

[19]. Latha R. Nair, David Peter S." Machine Translation Systems for Indian Languages" International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012

[20]. G.V. Garje, G.K. Kharate, "Survey of Machine Translation Systems in India", International Journal on Natural Language Computing (IJNLC), October 2013, Vol. 2, No.4, pp. 47-67 Available: http:// airccse.org/journal/ijnlc/current2013.html

[21]. Shachi Dave, Jignashu Parikh & Pushpak Bhattacharyya, (2002) "Interlingua-based English-Hindi Machine Translation and Language Divergence", Journal of Machine Translation , pp. 251-304.

[22]. Cristina España-Bonet, Gorka Labaka, Arantza Díaz de Ilarraza, Lluís Màrquez and Kepa Sarasola, (2011) "Hybrid Machine Translation Guided by a Rule–Based System", in proceedings of the 13 th Machine Translation Summit, September 19-23, 2011, Xiamen, China

[23]. Gorka Labaka, Cristina Espana-Bonet, Lluıs M`arquez, Kepa Sarasola, (2014) "A Hybrid Machine Translation Architecture Guided by Syntax", Journal of Machine Translation, pp. 1-29, Available at Springer via http://dx.doi.org/10.1007/s10590-014-9153-0