IJCSC 0973-7391 Volume 8 • Issue 1 Sept 2016 - March 2017 pp. 15-22 available online at www.csjournalss.com

Relevance of Statistical approach in todays era in Machine Translation

[1]Manu Raj Moudgil, [2]Dr.Preeti Dubey
[1]Research Scholar, JJT University, Jhunjhunu
[2]Assistant Professor,Department of Comp. Sc., Central University of Jammu, Jammu

Abstract: The research in the field of machine translation is speedup in the last few years. There are number of techniques are available for developing a machine translation system which includes word-word translation, syntactic transfer, Interlingua approach, Example based translation and statistical machine translation. Statistical machine translation is the latest approach of MT using which a lot of machine translations systems have been developed that have produced excellent results. This paper presents a brief overview of the systems developed using this approach and compares SMT systems with other systems developed using other approaches. The impact of SMT systems over the other approaches is presented in this paper. It also focuses on the idea that statistical machine translation method is optimum to increase the accuracy of Machine translation, as it removes the ambiguity, incompleteness and inconsistency in the sentences.

Keywords: Statistical machine translation, Decoding, Translation.

I. MACHINE TRANSLATION

Machine Translation (MT) is a easy way to automatically translating one natural language that is source language into another natural language that is target language. It should not be confused with computer-aided translation, machine-aided human translation. It is a sub field of computational linguistic that investigates the use of software to translate text from one language to other language MT was the first computer-based application related to natural language processing [1].

How MT is useful in today's Era

India is the developing country and we need to keep pace with fast modern technology and with present scenario of competition. For this it's necessary for us to create a strong bond with in our country as well as with other countries regarding technology, business, and education and in many other fields. Here the language can be the main hurdle which one can face in interaction with others. We can hope that a common language could resolve global problems that create conflicts. MT is the solution to this problem as this system helps in translating one language to another languages. India is a democratic country and having diversity of languages. There are more than 780 languages in India and machine translation is proved to be very useful in today's era for removing language barriers and make interaction easier. Today when we surfing internet, we come across many languages and characters which we don't understand, then we need some translation system to get familiar with that language or character. Now day's translation is used for administrative reports, instruction manual and many other documents. All this make the demand for translation increasing more quickly than the capacity of translators.

In coming time Machine translation will be even more popular—and necessary. According to industry experts, 15 million gigabytes of new content are generated every day. This volume is growing exponentially and is expected to increase by 20 times before 2020. By 2015 alone, there will be more than 15 billion devices connected to the Internet. All of these devices create a demand for access to more content in real-time, and the best way for global companies to meet these growing needs on an international scale may be through machine translation.

The future of MT is bright as it looks like it will continue to progress as companies seek to do more and more translation with their budgets.

II. *VARIOUS SYSTEMS BASED ON MT APPROACHES*

A lot of researches and work have been done on machine translation by time to time with the aim of translating one source language to another language, which can also be called target language, with more accuracy. Natural languages are quite complex, which make the machine translation a difficult task. Sometime many words have multiple meaning, sentences may have various reading and certain grammatical relation of one language might not exist in other language, and to face these challenges different systems are developed and taken into account.

Classification of MT system:

On the basis of core methodology of MT system, it is classified into two main paradigm i.e. Rule based approach and corpus based approach. In rule based approach human expert specify the set of rules for the translation. On the other hand, under the corpus-based approach the knowledge is automatically extracted by analysing translation examples from a parallel corpus built by human experts. The advantage is that, once the required techniques have been developed for a given language pair.[2] The systems developed under these approaches are discussed further.

A. Rule-based Machine Translation

Rule-based MT system use combination of language, rammer rule and dictionary words. It has much to do with the morphological, syntactic and semantic information about the source and target language. For Indian languages angla bharti and anu bharti is rule-based MT from English to Hindi and other Indian languages. The rule-based approach can further be divided into various approaches as shown in figure 1 below-

1. Direct Approach
2. Indirect approach or Transfer approach
3. Interlingua approach

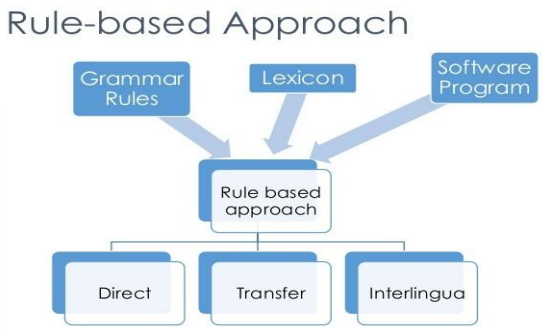

Fig:1 various methods of rule-based machine translation

1. Systems of Direct approach Machine Translation:

Direct approach address in number of ways, is the first generation MTS. In this approach the translation is based on large dictionaries and word by word translation. Word of source language are translated without passing through an additional/intermediary representation [3][4].

- ANUSAARKA (1995) is also a based on direct approach developed in Hyderabad covers all major Indian languages. It broken the MT system into two modules. The core anusaaraka output is in a language close to the target language, and can be understood by the human reader after some training. There is more accuracy in Indian languages as they share vocabulary and grammatical construction but face more ambiguity in different language as hindi to english, which are not closely related.[5]
- Gurpreet singh Joshan & gurpreet singh Lehal (2007) developed Punjabi to machine translation system using word to word translation and it has given accuracy above 92.8%. Due to the character gap in Punjabi and other languages, the word accuracy rate for the baseline is low.
- SYSTRAN ia also direct MTS hindi to punjabi and punjabi to hindi MTS developed in punjabi university patiala is based on direct translation approch.[6]
- Preety Dubey ,Devanand(2013) from jammu university ,have done work on translation of Hindi to Dogri language by using rule based approach and got above 95% results[7]. Results came out to be good enough as both are closely related languages using Devnagari script.

Above discussed systems shows that more the language is closely related to other , give good accuracy in results as compared to translation in foreign language, Other problems which these approaches faced were grammatical problem, and some time word of one language have different senses in other language which create ambiguity in translation.[7].

In statistical approach due to large corpus and ability to updation there is more accuracy, as it also able to make new pairs depending on the probability of occurrence. There is no linguistic knowledge required and its independent from pair of languages.

2 In-direct approach Machine Translation:

This approach is also called Transfer approach of machine translation. This approach to some extent is similar to Interlingua approach known as second generation of MT system from mid-1960's to 1980. It creates translation from an intermediate representation that simulates the meaning of original sentence.

This approach takes three steps of translation of text that is:

a) Analysis
b) Transfer
c) Generation

Firstly it analysis the source language in transformed into an abstract, less language specific representation. Secondly it transfer the syntactic/semantic structure of source language into the structure of target language. Then finally it generates the target language using billingual dictionaries and grammar rules [3][4].Some systems based on the transfer approaches are:

- RUSLAN (1985) developed for the translation of Chzech-russian language by charles university, pargue with research institute of mathematical machine. The result came out to be 40%correct translation, 40% post editing and 20% retranslation/editing. The ambiguities were left unresolved as there is no deep analysis of input sentences and main dictionary was also incomplete.[8]
- MANTRA (1999) is also a indirect approach for Indian languages funded by Government of India and the parser used for the language process is vyakarta[9-12]. This approach is general, however the lexicon/ grammar has been restricted to the sub language domain.

- English-to-Filipino MT system (2000) is a transfer based MT System that is designed and implemented using the lexical functional grammar (LFG) as its formalism. It involves morphological and syntactical analyses, transfer and generation stages. The whole

translation process involves only one sentence at a time.[8]

- Tamil-Hindi Machine-Aided Translation system (2009) L, Pralayankar P and Kavitha V developed a system which is based on Anusaaraka (started in 1984). MT system architecture is developed by Prof. C N Krishnan. It uses a lexical-level translation and has 80-85% coverage. Both stand-alone and web-based on-line versions have been developed. Tamil morphological analyzer and Tamil-Hindi bilingual dictionary (36K) are the biproducts of this system. It performs exhaustive syntactical analysis. They have also developed a prototype of English-Tamil MAT system. Currently, it has limited vocabulary (100-150 sentences) and small set of Transfer rules.

The major problem with the indirect approach is that the rules are must be used at the every step of machine translation. There are rules for syntactic/semantic (that is source language analysis rules) for source to target transfer and also the rules for the target language generation. Which make it more complex in functioning. On the other hand statistical MT system is easy to build and easy to maintain.

3 Interligua approach Machine Translation:

Third generation of machine translation is Interlingua approach. It developed to create linguistic homogeneity across the world. It is combination of two latin words that is inter and lingua where inter means between/intermediary and lingua means language. It is one instance of rule based MT approach. In this MT system translation is done initially by representing the source language text it an intermediary, known as Interlingua, before the translation of the interlingual language to the target language. Interlingual language is a neutral language which means it is independent of any language. The target language which resulted out from interlingua is also known as auxillary representation [3][9].

- CETA (1961), based on Interlingua and transfer-based approach, for translating Russian into French, it was developed at Grenoble University in France. It employed dependency-structure analysis of each sentence at the grammatical level and transfer mapping from one language-specific meaning representation at the lexical level. During the period of 1967-71, this system was used to translate about 4,00,000 words of Russian mathematics and physics texts into French. It was found that it fails for those sentences for which complete analysis cannot be derived.

At the end of 1990, institute of advance studies od united nations university Tokyo began multinational interlingua based MT project of universal networking language (UNN) which is based on standardized intermediary language. The Angla Bharti system developed at IIT Kanpur is also based on this approach [10-12].

- ANGLABHARTI (2001) R M K Sinha, Jain R, Jain A developed a machine aided translation system designed for translating English to Indian languages. It is developed using pseudo-interlingua approach. The International Journal on Natural Language Computing (IJNLC) Vol. 2, No.4, October 2013 54 interlingua approach made it possible to use the same system for translating English to more than one Indian language and has eliminated the need of developing separate translation system for English to each Indian language. The analysis of English as a source language is done only once and it creates intermediate structure – PLIL (Pseudo Lingua for Indian Languages). The PLIL is then converted to each Indian language through a process of text-generation. The effort for PLIL generation is 70% and text generation is 30%. Only with an additional 30% effort, new English to Indian language translation system can be built. The attempt has been made whereby has to do 90% translation task and remaining 10% is left for the human post-editing. The domain of this machine translation system has been public health.[13]
- The KANT system by Nyberg and MITAMURA in 1992 is created to translate caterpillar technical English (CTE) into other language. KANT's vocabulary(non–domain) specific is limited to a basic vocabulary of about 14000 distinct words sense while domain-specific technical terms are limited to a pre-defined vocabulary approx. 60000words and phrases for heavy equipment material. Structural restriction on the other hand, attempt to the limit the use of construction that would create difficulties in parsing such as the use of relative clauses.

The major disadvantage is the difficulties of defining an Interlingua, even for closely related languages (e.g. the Romance languages: French. Italian, Spanish, Portuguese). It is also very difficult to find out meaning from the texts in the original languages to produce the intermediate representation. Therefore the problem with the Interlingua approach is that it requires n analysis components Interlingua converters, and n generation components where n is the number of languages in Translation system. [7]

Which make this approach very complex and whereas statistical approach is less complex and having no ambiguity.

B. Corpus Based Machine Translation

In 1989 corpus based translation approach came forward as one of the most widely explored area in machine translation. It is also known as data-driven machine translation. This approach is proved as alternative approach for MT to overcome the problem of acquisition the problem of rule based MT. The level of accuracy of this approach make it dominant over the other

approaches. Corpus based machine translation uses bilingual parallel corpus [14]. This uses large number of raw data containing text and their translation. Following are the different types of corpus based machine translation:

1. Knowledge based Machine Translation
2. Example based Machine Translation
3. Statistical based Machine Translation

1. Knowledge based Machine Translation:

This system has more emphasis on complete understanding of the source text before translation into the target text. It is based on interlingual approach but it analysis the source language with more depth in comparison with interlingual approach. KVMT must be supported by world knowledge and by linguistic semantic knowledge about meaning of words and their combinations. The main advantage of this approach is it provide high quality of translation.

English-Vientnamese machine translation system is an example of KVMT. The KANT is also an example of KVMT. The KANTOO project is an object oriented C++ implementation of KANT technology for MT. LUTE project at NTT and ETL research, a Japanese multi-lingual project has also applied knowledge based approach [9][14].

This approach has increased accuracy of translation and efficient support of multiple target languages. KBMT systems provide high quality translations But there is need of lot of knowledge and efforts for the deep analysis of language or text.

2 .Example based Machine Translation:

This approach is also known as memory based translation. This approach is first developed by Makotonagao in 1984. It is based on recalling/finding analogous example of language pair or we can say it keep on using again and again the example of existing translation as the basis for new translation. It uses bilingual corpus with parallel text which means there are set of sentences in the source language and corresponding translation of each sentence in the target language with point to point mapping. There is process to extract and select equivalent phrases or word group from parallel bilingual text. The main focus of this approach is that if the previously translated sentence occurs again the same translation is likely to be correct again.

The process broken down into three stages: Matching, Alignment and recombination. In matching, the system finds the examples on the basis of their similarity with the input and then, it identify the paths of corresponding translation which are reused, is known as alignment. Finally in recombination identified paths of the example are put together in legitimate way. Angla bharti and Angla hindi are examples of EVMT [3][9].

In this approach there is need to move to a full-fledged EBMT environment, which means working with a real bilingual archive. The immediate problem to be solved then is the problem of alignment of the archive. If the full-sentence comparison method is used, it is sufficient to have the archive aligned at sentence level. If, however, the partitioning method is used, it becomes necessary to obtain alignment at the sub-sentential level. This latter task is, in fact, exactly the goal of the full-fledged statistical MT approaches. Results in text alignment have been achieved at IBM (e.g. Brown et al., 1990) and AT&T (e.g., Gale and Church, 1991) [15][16]. In the short run, the quality of sub-sentential alignment does not promise high-enough fidelity to support EBMT in a stable fashion. Because of this (and unconditionally for the full-sentence comparison method) a practical EBMT environment will have to involve a user interface, similar to the CMU TWS, to allow the human user to correct system output [17].

- ANUBAAD(2000,2004) Bandyopadhyay S developed a MT system which translates news headlines from English to Bengali using example based Machine Translation approach. An English news headline given to the system as an input is initially searched in the direct example-base for an exact match. If a match is found, the Bengali headline from the example-base is produced as output. If match is not found, the headline is tagged and the tagged headline is searched in the Generalized Tagged example-base. If a match is found in Generalized Tagged Example-Base, the Bengali headline is to be generated after appropriate synthesis. If a match is not found, the Phrasal example-base will be used to generate the target translation. If the headline still cannot be translated, the heuristic translation strategy is applied where translation of the individual words or terms in their order of appearance in the input headline will be generated. Appropriate dictionaries have been consulted for translation of the news headlines.[18]
- 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[19]

3. Statistical based Machine Translation:

This system uses statistical method on bilingual text corpus for generating translation. In 1949 Warren Werver was first who proposed the idea of statistical MT. But this method was adopted in 1950's and 1960's. This SBMT approach instead of using linguistic data it uses the statistical method such as n-gram based SMT, occurrence based SMT. The very first model of SMT was based on bayes theorem, which was proposed by brown et al. In this system it takes the view of every sentence in one language is a possible translation of any sentence in other

and the translation with the highest probability is considered to be the appropriate translation by the system. In simple language we can say the statistical based MT results by fetching those words from the given surrounding words which have highest level of probability of having its present position. All this processing needs a bilingual text corpus to create statistical rules which are based on probability of correct translation of a given word, phrase or sentence of the language.

- RAND Corporation in 1950-60 under took statistical analysis of large corpus of Russian physics texts, to extract bilingual glossaries and grammatical information. The Indian research lab IBM at New Delhi has initiated work on SMT between English and Indian languages. CANDIDE was the first SMT software given by IBM. Google language translators also uses statistical approach of translation [20][21].
- English to Indian Languages MT System (E-ILMT) (2006): The primary objective is to initially build an English-Hindi translation system capable of translation of free flow text as found on the web and gradually adapt it to other Indian language pairs as well. The training corpus (translation model) consisted of 5000 sentences and 800 sentences were split for testing and tuning. The baseline techniques used in this system were inadequate in producing a good quality translation. Therefore, pre-processing stage was included in the system which takes care of syntactic re-ordering on the source language to reduce long distance movements through SMT. It has helped to obtain a better phrase alignment table which resulted in a good improvement in the translation quality using Moses decoder with Giza++ alignment tool. The corpus (translation model) training size for achieving this effort was 12299 sentences with additional 1570 sentences split for testing and tuning. Some degradation in the output even after the syntactic processing was observed due to unavailability of sufficient corpus. The syntactically processed corpus was morphologically processed and used for training to counteract the problem of degradation in translation quality. A rule based suffix separation approach was used to separate the root word and the affixes due to the unavailability of sophisticated morphological analysers. The system is extended and tested for English-Marathi and English-Bengali pairs with the statistics.[13]
- HINDI to PUNJABI MT system (2015) Ajit Kumar in Punjabi university: Phrase-based Hindi-Punjabi machine translation system has been used to translate Hindi text to Punjabi. The quality of translated text is evaluated manually as well as automatically on BLEU score. The SPES has been applied to post edit the results obtained from the translation system and again evaluated extrinsically with the help of language experts. The improvement in the translation quality is evident from the results. But some errors are still present in the post-edited sentence. The reason behind this might be that Hindi words □□□□ (āpakī) and आप (āpa) are more likely translated as □□□□□□ (tuhām& ī) and □□□□□ (tusīm&) in the corpus and otherwise also. So the words which are translated differently in different context are potential source of error in the post-edited text also, and need to be handled separately. This is because; in one context the frequency of translation is higher as compared to other context, and such words retain their translation corresponding to high frequency context. In spite of this, it has been observed that most of the general grammatical errors are get corrected by the SPES trained on manually corrected corpus, as is evident from the extrinsic and intrinsic evaluation.[22][23]

C. Hybrid Approach Machine Translation:

On moving forward by keeping in mind the aim of more accuracy of translation or to overcome the limitations of other approaches has developed by the researchers. Hybrid approach is developed by taking the advantages of both statistical and rule-based translation. This approach proven to have better efficiency in the area of MT systems, It can be used in different ways .In some case rule based approach is used first for translation and then adjusting or correcting of output is done using statistical information. In other words we can say rules are used to pre-process the input data as well as post-process the statistical output of SBT system based translation.
This approach has more power flexibility and control in translation.

- Bengali to Hindi MT System (2009) Chatterji S, Roy D, Sarkar S and Basu A developed a hybrid Machine Translation system. It uses an integration of SMT with a lexical transfer based system (RBMT) i.e. multi-engine Machine Translation approach. The experimentation shows that BLEU scores of SMT and lexical transfer based system when evaluated separately are 0.1745 and .0424 respectively. The performance of hybrid system is better and its BLEU score is 0.2275 [24].
- 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 storing 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.

- The METIS II system is an example of hybridization around EVMT framework by Dirix et al 2005. TransEasy is a machine translation system based on hybrid approach. SisHiTra developed by Gonzalez et al is also hybrid MT system from Spanish to Catalan. This project combined knowledge based and corpus based techniques to produce a Spanish to Catalan MT system with no semantic constraints. Bengali to Hindi MT system developed at IIT kharagpur is also used hybrid approach of machine translation [24][25].

III. RESEARCH PROPOSAL

The objective of this research work is to determine that the statistical machine translation method is optimum when it is used with rule based MT in hybrid approach. Today in India only few SMT based systems are available but all those systems having much accuracy and are unambiguous. The next section of paper describes the comparison and drawbacks of the approaches towards the statistical based systems which is the present demand for all the linguists.

IV. DIFFERENT SYSTEMS WITH THEIR SHORTCOMINGS

A. Shortcomings of Rule based MT approach:

The following are the drawbacks which are related to the rule based machine translation system:

- Very less amount of good dictionaries are available and building a new dictionary is very costlier.
- Some linguistic data still has to be set Manually.
- It's exhausting to manage rule interactions in huge systems,ambiguity,and idiomatical expressi ons.
- Failure to adapt to new domains though RBMT systems typically give a mechanism to form new rules and extend and adapt the lexicon changes areusually very expensive and also the results, frequently, do not pay off.

a) Shortcomings of Direct based MT approach:

The following are the drawbacks which are related to the rule based machine translation system:

- The limitation of this approach is apparent. It are often characterised as 'word-for-word ‘translation with some native word-order adjustment. It gave the sort of translation quality that may be expected from somebody with a very low cost lexicon and solely the foremost rudimentary information of the descriptive linguistics of the target language: Frequent mistranslations at the lexical level and mostly inappropriate syntax structures that reflected too closely those of the source language.
- Lack of linguistic and commutative knowledge is also one of the issue. From the linguistic point of view there is no analysis of internal structure of source text grammatically. There is also lack of computational sophistication was largely a reflection the primitive state of computer sciences.

b) Shortcomings of Indirect or Transfer based MT approach:

The following are the shortcomings which are related to the indirect or transfer based machine translation system:

- The major problem with the indirect approach is that the rules are must be used at the every step of machine translation. There are rules for syntactic/semantic (that is source language analysis rules) for source to target transfer and also the rules for the target language generation.
- It is also difficult to do a lot of work in reusable modules of analysis and synthesis.
- It is very difficult to maintain the transfer modules very simple.

c) Shortcomings of Interlingua based MT approach:

The following are the shortcomings which are related to Interlingua based machine translation system:

- A lot of problems in defining the Interlingua, even for the closely related languages for example French, Spanish, Italian, Portuguese. From the past decades a truly universal Interlingua and language independent Interlingua is defined as then best efforts of linguists.
- It is also very difficult to find out meaning from the texts in the original languages to produce the intermediate representation.
- Semantic differentiation is target language specific and making such distinctions are comparable to lexical transfer and not all distinctions needed for the translation.

B. Shortcomings of Corpus based MT approach:

The short comings of corpus based MT is divided into three categories knowledge based MT, Example based MT and Statistical Based MT.

***a)** Shortcomings of knowledge based MT approach:*

The following are the drawbacks which are related to the knowledge based machine translation system:

- A lot of effort is required to build up the knowledge bases.
- An operative definition of the size of the knowledge base
- Also the choice of the representation language with its required logical or formal properties.

- It is time consuming and labour intensive, as there is need of large amount of hand coded lexical knowledge.

b) *Shortcomings of Example based MT approach:*

The following are the drawbacks which are related to the example based machine translation system:

- Example based machine translation is also very popular way of translation, because there is no need of manual derived rules like syntactic/semantic rules as in transfer approach. But it needs analysis and generation modules to supply the dependency trees required for the examples information and for analysing the sentence.
- Second important problem with example based machine translation is the computational efficiency mainly for the large databases or systems although the parallel computation techniques may be applied.

c) *Shortcomings of Statistical based MT approach:*

The following are the shortcomings which are related to the statistical based machine translation system:

- The major problem of SMT is to developing the corpus as it required a lot of human effort and cost.
- Statistical machine translation gives good results when applied to the closely related languages but it not work properly, when it applied to the languages that having significantly different word orders.
- Their benefits gives more importance to European languages.

C. Shortcomings of Hybrid based MT approach:

The following are the shortcomings which are related to the statistical based machine translation system:

- As hybrid approach successfully collaborate the benefits of all approaches it also have the some common limitations.
- It maintains the expensiveness of rule-based machine translation as it purposes additional complexities of maintain side-by-side systems creating their right commercial value questionable.

V. RESULT AND DISCUSSION BASED ON COMPARISON OF APPROCHES WITH SMT

We have discussed the working of all approaches of machine translation, how each and every approach having process of translation of one source language to target language. On the basis of above discussion we can say that every approach has its advantages and limitations. Direct approach, transfer approach and inter-lingua approach are the types of rule based machine translation system. RBMT system is developed on the basis of morphological, syntactic and semantic analysis of both source and target languages and it belongs to domain of rationalism whereas systematic machine translation which is based on corpus based machine translation system is generated on the analysis of bilingual text of corpora and it belongs to empiricism. Rule based methods emphasized to understand the grammar rules on the other hand statistical approach ignores or pay very less attention to grammar of a particular language. Rule based methods are based on human knowledge as human uses their knowledge and experience to prepare rules. So it is necessary to have deep knowledge of languages. In spite of that it is very hard to capture all rules. Moreover there is great difficulty of correction of input or add new rules to the system to generate the translation in rules based methods. In contrast, adding more examples to statistical machine translation can improve the system as it based on the data. No matter what the shortcoming rule based approaches have, it is still valuable approach for machine translation from the syntactic point of view. However the attributes which make SMT more advantageous over other approaches are:

Firstly it uses bitext as the fundamental source of data. Secondly, it is experimental which is based on machine learning instead of rational knowledge with linguistic writing rules. Thirdly, it can be improved easily by the getting more data. Fourthly, it can develop new language pair by finding suitable parallel corpus data. Overall these attributes the main trait of SMT which is not shared by other corpus based machine translation system is , it uses statistical data such as parameters and the probabilities derived from bitext, in which processing data is essential and even if the input is in the training data, the same translation is not guaranteed.

By keeping in view of getting more accuracy in translation system, researchers have combined both rule based machine translation systems and statistical methods to develop the new approach which is called hybrid approach. But SMT is main base in those system and plays a vital role in developing the hybrid systems. Furthermore some additional important advantages of SMT systems are fast development cycles, robust-ness, very superior lexical selection and fluent due to use of different language models.

VI. CONCLUSION

This paper concludes that how the statistical machine translation system is optimum with the collaboration of rule based system called hybrid systems rather than the other approaches of machine translation From the above comparison of results and discussions of all the approaches with their systems we can say that the accuracy level of the translated text is much increased if

we used the hybrid approach which includes the Statistical based MT system to increase the accuracy and by removing all ambiguities, incompleteness and inconsistencies.

VII. ACKNOWLEDGEMENT

My sincere thanks to Dr. Vishal Goyal, Assistant Professor, Department of Computer Science, Punjabi University, for his guidance.

VIII. REFERENCES

[1] WJohn Hutchins and Harold L Somers, (1992), "An introduction to machine translation" London: Academic Press, [ISBN: 0-12-362830-X]

[2] Marta R. Costa-Jussa, Mireia Farrus, Jos´e B. Marino˜ , Jos´e A. R. Fonollosa, "Study And Comparison of Rule-Based And Statistical Catalan-Spanish Machine Translation Systems", Computing and Informatics, Vol. 31, 2012, 245–270

[3] Tripathi, Sneha; Sarkhel, Juran Krishna; (2010). "Approaches to machine translation", Annals of Library and Information Studies Vol. 57, December 2010, pp. 388393.

[4] Hutchins, W. John, Somers, Harold L. An Introduction to Machine Translation. AcademicPress, London, 1992.

[5] Akshar Bharati, Vineet Chaitanya, Amba P. Kulkarni and Rajeev Sanga," ANUSAARAKA: MACHINE TRANSLATION IN STAGES", IIT Kanpur Centre for NLP at Hyderabad Central University Campus Hyderabad, A Quarterly in Artificial Intelligence, Vol.10, No.3 (July 1997), NCST, Mumbai, pp.22-25

[6] Vishal Goyal, "Development of a Hindi to Punjabi Machine Translation System", PhD Thesis, Department of Computer Science, Punjabi University, Patiala, 2010.

[7] Preeti Dubey, Devanand," Machine Translation System for Hindi-Dogri Language Pair", in proceedings of IEEE Conference (ICMIRA), held at SMVDU, in Dec 2013, ISBN: 978-0-7695-5013-8, pages: 422-425, DOI 10.1109/icmira.2013.89

[8] H. B. Sale, 2Dr. N. K. Rana," Evaluation of features & Comparative Study of Machine Translation Systems in Non-Indian Languages", IJCST Vol. 3, Issue 2, April - June 2012, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

[9] HAROLD SOMERS, Review Article: Example-based MachineTranslation , Machine Translation 14: 113–157, 1999.

[10] Sitender, Seema Bawa, "Survey of Indian Machine Translation Systems", IJCST Vol. 3, Issue 1, Jan. - March 2012, ISSN : 0976-8491 (Online)

[11] Salil Badodekar , "Translation Resources, Services and Tools for Indian Languages", Available Online: http://www.cfilt.iitb.ac.in/Translation-survey/survey.pdf

[12] Sneha tripathi, et.al, "Approaches to Machine Translation", Annals of Library & Information Studies, vol 57, dec 2010, pp: 388-393

[13] G V Garje1 and G K Kharate," SURVEY OF MACHINE TRANSLATION SYSTEMS IN INDIA", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.4, October2013

[14] Nyberg, Eric; Mitamura, Teruko; and Carbonell, Jaime G., "The KANT Machine Translation System: From R&D to Initial Deployment" (1997). Computer Science Department.Paper339.http://repository.cmu.edu/compsci/339

[15] Brown. P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R.L. and Roossin P.S. 1990. A statistical approach to language translation, Computational Linguistics, vol 16, 79-85.

[16] Gale, W. and K. Church. 1991. Identifying word correspondence in parallel text. Proceedings of the DARPA NLP Workshop.

[17] Sergei Nirenburg, Constantine Domashnev and Dean J. Grannes, Two Approaches to Matching in Example-Based Machine Translation, School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 USA

[18] S. Bandyopadhyay, (2004) "ANUBAAD - The Translator from English to Indian Languages", in proceedings of the VIIth State Science and Technology Congress. Calcutta. India. pp. 43-51

[19] R. Mahesh K. Sinha & Anil Thakur, (2005) "Machine Translation of Bi-lingual Hindi-English (Hinglish) Text", in proceedings of 10th Machine Translation Summit organized by Asia-Pacific Association for Machine Translation (AAMT), Phuket, Thailand

[20] Ruslan Mitkov, "The Oxford Handbook of Computational Linguistics", Oxford University Press, 2003, chapter 28, pages: 513-528, ISBN: 0-19-823882-7

[21] http://en.wikipedia.org/wiki/Machine_translation

[22] Ajit Kumar and Vishal Goyal, (2011) "Comparative Analysis of Tools Available for Developing Statistical Approach Based Machine Translation System", ICSIL 2011 CCIS 139, pp 254-260

[23] Ajit Kumar and Vishal Goyal," Statistical Post Editing System (SPES) Applied to Hindi-Punjabi PB-SMT System", Indian Journal of Science and Technology.Vol8(27),DOI:10.17485/ijst/2015/v8i27/82463, October 2015 ISSN (Print) : 0974-6846 ISSN (Online) : 0974-5645.

[24] Sanjay Chatterji et.al, "A Hybrid Approach for Bengali to Hindi Machine Translation", Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, pages: 81-91. Also Available at: http://ltrc.iiit.ac.in/proceedings/ICON-2009

[25] J.González, A.L.Lagarda, J.R.Navarro, L.Eliodoro, A.Giménez,F.Casacuberta, J.M.de Val, & F.Fabregat. 2006. SisHiTra: a Spanish-to-Catalan hybrid Machine Translation system. LREC- 2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages:"Strategies for developing Machine Translation for minority languages", Genoa, Italy, 23 May 2006. pp. 69-73