IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Constructing Challenge Set For Machine Translation

Ms.Saumya Jain, saumyajain773@gmail.com

Teerthanker Mahaveer University

Dr. Ranjana Sharma, ranjana.computers@tmu.ac.in

Teerthanker Mahaveer University

Mr. Ajay Rastogi, ajay.computers@tmu.ac.in

Teerthanker Mahaveer University

Abstract

In the modern world, there is an increased need for language translations owing to the fact that language is an effective medium of communication. The demand for translation hasbecome more in recent years due to increase in the exchange of information between various regions using different regional languages. Accessibility to web document in other languages, for instance, has been a concern for information Professionals. Machine translation (MT), a subfield under Artificial Intelligence, is the application of computers to the task of translating texts from one natural (human) language to another. Many approaches have been used in the recent times to developan MT system. Each of these approaches has its own advantages and challenges. This paper takes a look at these approaches with the few of identifying their individual features,challenges and the best domain they are best suited to.

Keywords: *Machine Translation, Rule-based Approach,Corpus-based Approach, Statistical Approach, Transfer-Based Approach*

1. Introduction

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

The idea of machine translation may be traced back to the 17th century. In 1629, René Descartes proposed a universal language, with equivalent ideas in different tongues sharing one symbol. The field of ―machine translation‖ appeared in WarrenWeaver‘s Memorandum on Translation (1949). The first researcher in the field, Yehosha Bar-Hillel, began his researchat MIT (1951). A Georgetown MT research team followed(1951) with a public demonstration of its system in 1954. MT research programmes popped up in Japan and Russia (1955), and the first MT conference was held in London (1956). Researchers continued to join the field as the Association for Machine Translation and Computational Linguistics was formed in the U.S. (1962) and the National Academy of Sciences formed the Automatic Language Processing Advisory Committee (ALPAC) to study MT (1964). Real progress was much slower, however, and after the ALPAC report (1966), which found that the ten-

year-long research had failed to fulfill expectations, funding was greatly reduced. The idea of using digital computers for translation of natural languages was proposed as early as 1946 by A. D. Booth and possibly others.

Machine Translation or MT or robotized interpretation is simply a procedure when a computer software translates text from one language to another without human contribution. At its fundamental level, machine translation performs a straightforward replacement of atomic words in a single characteristic language for words in another.

1.1 Translation process

To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restoredin the target language, i.e., the translation. While on the surface, this seems straightforward, it is far more complex,

The human translation process, for instance, may be described as:

1. Decoding the meaning of the source text; and
2. Re-encoding this meaning in the target language.

Behind this ostensibly simple procedure lies a complex cognitive operation. To decode the meaning of the source text in its entirety, the translator must interpret and analyse all the features of the text, a process that requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc., of the source language, as well as the culture of its speakers. The translator needs the same in-depth knowledge to re-encode the meaning in the target language.

Since natural languages are highly complex, MT becomes a difficult task. Many words have multiple meanings, sentences may have various readings, and certain grammatical relations inone language might not exist in another language. The following diagram shows all the phases involved in the process of Machine Translation.

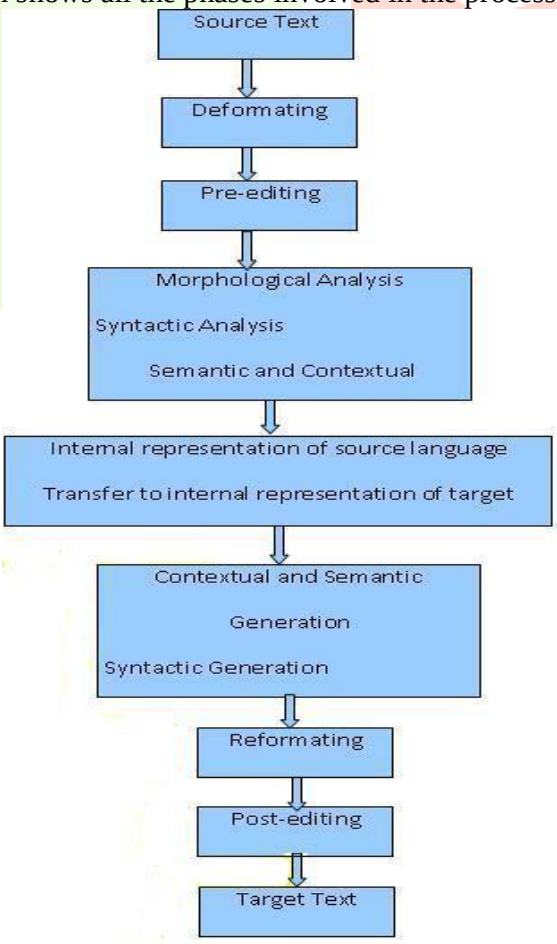


Fig. 1 A Typical Machine Translation Process(source: worldofcomputing.net)

A machine translation (MT) system first analyses the source language input and creates an internal representation. This representation is manipulated and transferred to a form suitable for the target language. Then at last output is generated in the target language. On a basic level, MT performs simple substitution of words in one natural language for words inanother, but that alone usually cannot produce a goodtranslation of a text because recognition of whole phrases and their closest counterparts in the target language is needed.

Therein lies the challenge in machine translation: how to program a computer that will "understand" a text as a person does, and that will "create" a new text in the target language that "sounds" as if it has been written by a person.

This problem may be approached in a number of ways. This paper takes a look at these approaches and their attendant challenges.

2. Literature Review

Sr No.	Author	Year Of Publication	Title	Work Done
1	Marta R. Costa-Jussà, Mireia Farrús, José B. Mariño	2012-07-18	Study and Comparison of Rule-Based and Statistical Catalan-Spanish Machine Translation Systems	This paper reports a study and comparison of several specific Catalan-Spanish machine translation systems: two rule-based and two corpus-based (particularly, statistical-based) systems, results show all the evaluations performed are characterised by some degree of correlation, and human evaluators tend to be specially critical with semantic and syntactic errors.
2	V. Laximi and H. Kaur	**2013**	A Survey of Machine Translation Approaches	In this paper a brief overview of the MT and various techniques of designing an MT system is being presented. Also the challenges upcoming while translating one language into another are also discussed.
3	Peter Dirix, Ineke Schuurman, and Vincent Vandeghinste	2015	METIS-II: Example-based machine translation using monolingual corpora System description	The METIS-II project1 is an example-based machine translation system, making use of minimal resources and tools for both source and target language, making use of a target-language(TL) corpus, but not of any parallel corpora. In the current paper, the discussion is the viewpoint of the team on the general philosophy and outline of the METIS-II system
4	L. Dugast, J. Senellart and P. Koehn	**2007**	Statistical Post-Editing on SYSTRAN‘s Rule-based Translation System	This paper describes the combination of a SYSTRAN system with a “statistical postediting” (SPE) system. We document qualitative analysis on two experiments performed in the shared task of the ACL 2007 Workshop on Statistical Machine Translation. Comparative results and more integrated “hybrid” techniques are discussed.

5	D. Groves and A.Way,	2005	Hybrid Example-based SMT: the Best of Both Worlds	This paper work provides an in depth comparison of the Example-Based Machine Translation (EBMT) system with a Statistical Machine Translation (SMT) system constructed from freely available tools. According to a wide variety of automatic evaluation metrics, the authors demonstrated that their EBMT system outperformed the SMT system by a factor of two to one.
6	P. Koehn and H. Hoang	2007 Future findings	Factored Translation Models	This paper present an extension of phrase-based statistical machine translation models that enables the straight-forward integration of additional annotation at the word-level — may it be linguistic markup or automatically generated word classes. In a number of experiments it has been shown that factored translation models lead to better translation performance, both in terms of automatic scores, as well as more grammatical coherence.
7	R. Harshawardhan	2011 future work	Rule-based Machine Translation System for English to Malayalam Language	A rule-based machine translation system for English to Malayalam language pair has been developed (Model) in this paper work. The machine translation system takes in the English sentence as input and parse with the help of Stanford Parser. The Stanford Parser is made use for four main purposes on the source (English) side processing, in the machine translation system: Parsing, POS tagging, Stemming and Morphological analysis. The English to Malayalam bilingual dictionary is created within this research paper.

3. Methodology

3.1 Machine Translation Approaches

A machine translation (MT) system first analyses the source language input and creates an internal representation. This representation is manipulated and transferred to a form suitable for the target language. Then at last output is generated in the target language.
MT systems can be classified according to their core methodology. Under this classification, two main paradigms can be found: the rule-based approach and the corpus-based approach. In the rule-based approach, human experts specify a set of rules to describe the translation process, so that an enormous amount of input from human experts is required. On the other hand, under the corpus-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Combining the features of the two major classifications of MT systems gave birth to the Hybrid Machine Translation Approach.

3.1.1 Rule-Based Machine Translation (RBMT)Approach

Rule-Based Machine Translation (RBMT), also known as Knowledge-Based Machine Translation and Classical

Approach of MT, is a general term that denotes machine translation systems based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. Having input sentences (in some source language), an RBMT system generates them to output sentences (in some target language) on the basis of morphological, syntactic, and semantic analysis of both thesource and the target languages involved in a concrete translation task.

3.1.1.1 Basic Principles of RBMT Approach

RBMT methodology applies a set of linguistic rules in three different phases: analysis, transfer and generation. Therefore, a rule-based system requires: syntax analysis, semantic analysis, syntax generation and semantic generation. Speaking in generalterms, RBMT generates the target text given a source textfollowing the steps shown in Fig. 2

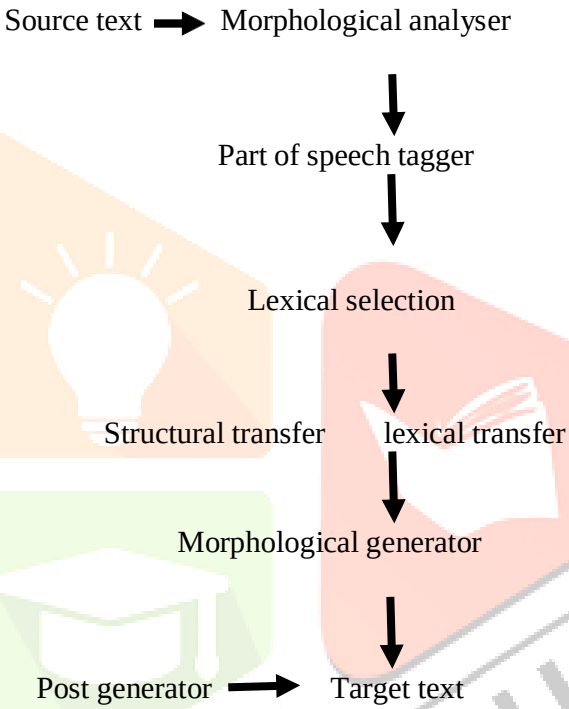


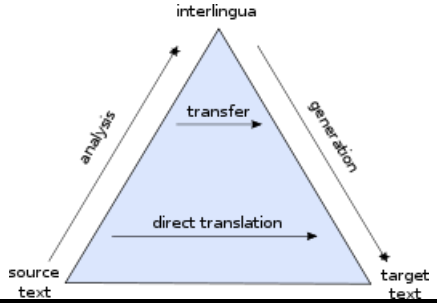
Fig. 2 Architecture of the RBMT Approach

The main approach of RBMT systems is based on linking the structure of the given input sentence with the structure of the demanded output sentence, necessarily preserving their unique meaning.

3.1.1.2 Issues of RBMT Approach

The following are the shortcomings that are associated withRBMT approach:

- Insufficient amount of really good dictionaries.Building new dictionaries is expensive.
- Some linguistic information still needs to be setmanually.
- It is hard to deal with rule interactions in big systems,ambiguity, and idiomatic expressions.
- Failure to adapt to new domains. Although RBMT systems usually provide a mechanism to create new rules and extend and adapt the lexicon, changes are usually very costly and the results, frequently, do



not pay off.

○ **Direct Machine Translation (DMT) Approach** Starting with the shallowest level at the bottom of the pyramidis the Direct Machine Translation Approach. DMT approach isthe oldest and less popular approach. Direct translation is madeat the word level. Machine translation systems that use thisapproach are capable of translating a language, called sourcelanguage (SL) directly to another language, called targetlanguage (TL). Words of the SL are translated without passingthrough an additional/intermediary representation. The analysisof SL texts is oriented to only one TL. Direct translationsystems are basically bilingual and uni-directional. Directtranslation approach needs only a little syntactic and semanticanalysis. SL analysis is oriented specifically to the productionof representations appropriate for one particular TL. DMT is aword-by-word translation approach with some simplegrammatical adjustments.

Challenges of a DMT System

- ❖ The limitation of this approach is obvious. It can be characterized as _word-for-word‘ translation with some local word-order adjustment. It gave the kind of translationquality that might be expected from someone with a very cheap bilingual dictionary and only the most rudimentary knowledge of the grammar of the target language: frequentmistranslations at the lexical level and largely inappropriate syntax structures which mirrored too closely those of the source language.
- ❖ The linguistic and computational naivety of this approachis also an issue. From a linguistic point of view what is missing is any analysis of the internal structure of thesource text, particularly the grammatical relationships between the principal parts of the sentences. The lack of computational sophistication was largely a reflection of theprimitive state of computer science at the time, but it was also determined by the unsophisticated approach to linguistics in MT projects of the late 1950s.

○ **Interlingual Machine Translation Approach**

The failure of the first generation systems led to the development of more sophisticated linguistic models for translation. In particular, there was increasing support for the analysis of source language texts into some kind of intermediate representation — a representation of its _meaning‘ in some respect — which could form the basis of generation of the target text. Interlingual machine translation is one instance of rule-based machine-translation approaches. In this approach,the source language, i.e. the text to be translated, is transformedinto an interlingual language, i.e. a –language neutral‖ representation that is independent of any language. The target language is then generated out of the interlingua. One of the major advantages of this system is that the interlingua becomes more valuable as the amount of target languages it can beturned into increases.

Challenges of Interlingual Machine Translation Approach

- ❖ There are the difficulties in defining an interlingua, evenfor closely related languages (e.g. the Romance languages: French. Italian, Spanish, Portuguese). A truly _universal‘ and language-independent interlingua has defied the best efforts of linguists over the years.
- ❖ It is difficult to extract meaning from texts in the original languages to create the intermediate representation.
- ❖ Semantic differentiation is target-language specific and making such distinctions is comparable to lexical transfer not all distinctions needed for translation

○ **Transfer-based Machine Translation Approach**

Because of the disadvantage of the Interlingua approach, a better rule-based translation approach was discovered, calledthe Transfer-based Approach. Transfer-based machine translation is similar to interlingual machine translation in thatit creates a translation from an intermediate representation thatsimulates the meaning of the original sentence. Unlikeinterlingual MT, it depends partially on the language pairinvolved in the translation. On the basis of the structuraldifferences between the source and target language, a transfer system can be broken down into three different stages: i)Analysis, ii) Transfer and iii) Generation. In the first stage, theSL parser is used to produce the syntactic representation of aSL sentence. In the next stage, the result of the first stage isconverted into equivalent TL-oriented representations. In thefinal step of this translation approach, a TL morphologicalanalyzer is used to generate the final TL texts. It is possible with this translation approach to obtain fairly high qualitytranslations, with accuracy in the region of of 90%.

Challenges of Transfer-based Machine Translation

- ❖ One of the problems with transfer Based Machine translation approach is that rules must be applied at every step of translation. There are rules for source language analysis (syntactic/semantic), rules for source-to-target transfer and rules for target language generation
- ❖ It is difficult to do as much work as possible in reusable modules of analysis and synthesis.
- ❖ It is difficult to keep transfer modules as simple as possible.

3.1.2 Corpus-based Machine TranslationApproach

Corpus based machine translation (also referred as data driven machine translation) is an alternative approach for machinetranslation to overcome the problem of knowledge acquisition problem of rule based machine translation. Corpus Based Machine Translation (CBMT) uses, as it name points, a bilingual parallel corpus to obtain knowledge for new incomingtranslation. This approach uses a large amount of raw data inthe form of parallel corpora. This raw data contains text and their translations. These corpora are used for acquiring translation knowledge. Corpus based approach is further classified into following two sub approaches: Statistical Machine Translation and Example-based Machine Translation Approach.

- **Statistical Machine Translation Approach**

Statistical machine translation (SMT) is generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The initial model of SMT, based on Bayes Theorem, proposed by Brown et al. takes the view that every sentence in one language is a possible translation of any sentence in the other and the most appropriate is the translation that is assigned the highest probability by the system. The idea behind SMT comes from information theory. A document is translated according to the probability distribution function indicated by p(e|f), which is the Probability of translating a sentence f in the SL F (for example, Hindi) to a sentence e in the TL E (for example, English). The problem of modeling the probability distribution p(e|f) has been approached in a number of ways. One intuitive approach is to apply Bayes theorem. That is, if p(f|e) and p(e) indicate translation model and language model, respectively, then the probability distribution $p(e|f) \propto p(f|e)p(e)$. The translation model p(f|e) is the probability that the source sentence is the translation of the target sentence or the way sentences in E get converted to sentences in F. The language model p(e) is the probability of seeing that TL string or thekind of sentences that are likely in the language E. This decomposition is attractive as it splits the problem into two sub problems.

The translation model ensures that the machine translation system produces target hypothesis corresponding to the source sentence. The language model ensures the grammatically correct output.

Issues with statistical machine translation include:

Sentence Alignment: In parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa. Sentence aligning can be performed through the Gale-Church alignment algorithm.

Statistical Anomalies: Real-world training sets may override translations of, say, proper nouns. An example would be that "I took the train to Berlin" gets mis-translated as "I took the train to Paris" due to an abundance of "train to Paris" in the training set.

Data Dilution: This is a common anomaly caused when attempting to construct a new statistical model (engine) to represent a distinct terminology (for a specific corporate brand or domain). Training sets used from alternative sources to the specific brand to compensate for a limited quantity of brand- specific corpora may ‗dilute' brand terminology, choice of words, text format and style.

Idioms: Depending on the corpora used, idioms may not translate "idiomatically".

Different word orders: Word order in languages differ. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages. There are also additional differences in word orders, for instance, where modifiers for nouns are

located, or where the same words are used as a question or a statement.

Challenges of Statistical Machine Translation Approach

- Corpus creation can be costly for users with limited resources.
- The results are unexpected. Superficial fluency can be deceiving.
- Statistical machine translation does not work well between languages that have significantly different word orders (e.g. Japanese and European languages).
- The benefits are overemphasized for European languages.

- **Example-based Machine Translation Approach**

Example-based machine translation (EBMT) is characterized by its use of bilingual corpus with parallel texts as its main knowledge, in which translation by analogy is the main idea. An EBMT system is given a set of sentences in the source language (from which one is translating) and corresponding translations of each sentence in the target language with point to point mapping. These examples are used to translate similar types of sentences of source language to the target language. There are four tasks in EBMT: example acquisition, example base and management, example application and synthesis. At the foundation of example-based machine translation is the idea of translation by analogy. The principle of translation by analogy is encoded to example-based machine translation through the example translations that are used to train such a system.

Example-based machine translation systems are trained from bilingual parallel corpora, which contain sentence pairs like the example shown in the table. Sentence pairs contain sentences in one language with their translations into another. The particular example shows an example of a minimal pair, meaning that the sentences vary by just one element. These sentences make it simple to learn translations of subsentential units.

Challenges of EBMT approach

EBMT is an attractive approach to translation because it avoids the need for manually derived rules. However, it requires analysis and generation modules to produce the dependency trees needed for the examples database and for analyzing the sentence. Another problem with EBMT is computational efficiency, especially for large databases, although parallel computation techniques can be applied.

3.1.3 Hybrid Machine Translation Approach

By taking the advantage of both statistical and rule-based translation methodologies, a new approach was developed, called hybrid-based approach, which has proven to have better efficiency in the area of MT systems. At present, several governmental and private based MT sectors use this hybrid- based approach to develop translation from source to target language, which is based on both rules and statistics. The hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. This technique is better than the previous and has more power, flexibility, and control in translation.

Hybrid approaches integrating more than one MT paradigm are receiving increasing attention

4. Results and Discussion

Each machine translation approach has its advantages and disadvantages. What one approach possesses, the other one seems to be lacking and vice versa. Rule-based methods focus on trying to understand the grammar rules while the statistical approach pays very minimal or no attention to the grammar of a particular language.

The rule-based machine translation approach has been implemented in computational linguistics since its very early days. Human involvement in this approach is significant as it is the human agent who creates the rules. In other words, the humans use their knowledge and experience to prepare the rules. The benefit of this approach is that it analyzes the input on the syntactic and - to an extent - semantic levels. The downside to rule-based machine translation is that it requires a deep linguistics knowledge as well as long time to prepare the rules. In the end, to capture all rules would be extremely hard. Nevertheless, the rule-based approach is very valuable for machine translation, especially from the syntactic point of view. Rule-based machine translation can be constantly modified as one can analyze the rules that do not produce a desirable output and thus focus on fixing

the problem. This approach can be a great starting point for those languages where a parallel bilingual corpus does not exist yet.

For Corpus-based approach the knowledge is automatically extracted by analysing translation examples from a parallelcorpus built by human experts. The advantage is that, once the required techniques have been developed for a given language pair, MT systems should – theoretically – be quickly developedfor new language pairs using provided training data. Adding more examples to a Corpus-based system can improve the system since it is based on the data, though the accumulation and management of the huge bilingual data corpus can also be costly.

Hybrid machine translation is a method of machine translation that is characterized by the use of multiple machine translation approaches within a single machine translation system. The motivation for developing hybrid machine translation systems stems from the failure of any single technique to achieve a satisfactory level of accuracy. Many hybrid machine translationsystems have been successful in improving the accuracy of the translations. Nowadays, the most widely used MT systems (Hybrid) use the rule-based and the statistical approaches.There have been several research works which combine both approaches.

5. Conclusion

Machine translation has been an active research subfield of artificial intelligence for years. Machine translation (MT) is a hard problem, because natural languages are highly complex, many words have various meanings and different possible translations, sentences might have various readings, and the relationships between linguistic entities are often vague. In addition, it is sometimes necessary to take world knowledge into account. The number of relevant dependencies is much too large and those dependencies are too complex to take them all into account in a machine translation system. Given these boundary conditions, a machine translation system has to make decisions (produce translations) given incomplete knowledge. This problem may be approached in a number of ways. This paper took a look at these approaches and their attendant challenges. The work shows that there is no perfect approach, though the problems associated with some of the approachesare very minimal. Combining some of the best features of some approaches to form a hybrid approach helps in taking care of the challenges posed by many approaches.

6. Future Scope

The robustness of MT needs further improvement. Sometimes, a slight change in the source sentence—such as a word or punctuation mark—can lead to great changes in the translation. However, human beings have a strong error-tolerant ability that allows them to flexibly deal with various non-standard language phenomena and errors, and sometimes even unconsciously correct them. Robust MT systems are crucial in real applications. Developing explainable MT methods may be one possible solution.

Also the NMT methods are facing serious data sparseness problems in resource-poor language pairs and domains. The current MT systems often use tens of millions or even hundreds of millions of sentence pairs of data for training. Otherwise, the translation quality will be poor. However, human beings can learn from only a small number of samples. Although many data-augmentation methods, multitask learning methods, and pretraining methods have been proposed to alleviate this problem, the question of how to improve the translation quality for resource-poor language pairs remains open.

In summary, there is still a long way to go to achieve high-quality MT. It is necessary to develop new methods that can combine symbolic rules, knowledge, and neural networks to further improve translation quality. Fortunately, the use of MT in real applications continues to provide more data, promoting the quick development of new MT methods.

7. References

[1] M.R. Costa-Jussa, M. Farrus, J.B. Marino and J.A. Fonollosa), ‖Study and Comparison of Rule-based and Statistical Catalan- Spanish Machine Translation Systems‖, Computing and Informatics, Vol. 31, 2011,pp 245-270

[2] V. Laximi and H. Kaur. ‖A Survey of Machine Translation Approaches‖, International Journal of Science, Engineering and Technology Research, Vol. 2, Issue 3, 2013, pp 716-719

[3] P. Dirix, I. Schuurman and V. Vandeghinste, ‖Metis II: Example-based Machine Translation using Monolingual Corpora - System Description‖, In Proceedings of the 2nd Workshop on Example-Based Machine Translation, 2005, pp 43-50.

[4] L. Dugast, J. Senellart and P. Koehn, ‖Statistical Post- Editing on SYSTRAN's Rule-based Translation System‖, In Proceedings of the Second Workshop on SMT, 2007, pp 220-223.

[5} D. Groves and A.Way, ‖Hybrid Example-based SMT: the Best of Both Worlds‖, In Proceedings of the ACL Workshop on Building and Using Parallel Texts, pp 183-190.

[6] P. Koehn and H. Hoang, ‖Factored Translation Models‖, In Proceedings of the 2007 Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning, 2007, pp, 868-876

[7] P. Brown and B. Alli, "A Statistical Approach to Machine Translation", Computational Linguistics, Vol. 16, No.2, 1990, pp 79-85.

[8] S. Tripathi and J.K. Sarkhel, ‖Approaches to Machine Translation‖, Annals of Library and Information Studies, Vol 57, 2010, pp 388-393

[9] P.J. Anthony, ―Machine Translation Approaches and Survey for Indian Languages‖, Computational Linguistics and Chinese Language Processing, Vol 18, No. 1, 2013, pp 47-78

[10] M.S. Henley, ‖The Use of Context-free Grammar in Support of Slovak-English Rule-based Machine Translation‖, M.Sc. Thesis, Faculty of Graduate School of Arts and Science, Georgetown University, Washington DC, 2008.

[11] R. Harshawardhan, ‖Rule-based Machine Translation System for English to Malayalan Language, M.Sc. Thesis, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbarore, 2011.

[12] A.A. Mohammed, ―Machine Translation of Noun Phrases: From English to Arabic, M.Sc. Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt, 2000.