(IJRST) 2019, Vol. No. 9, Issue No. IV, Oct-Dec **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

EMPLOYABILITY OF MACHINE LEARNING TOOLS AND TECHNIQUES TO TRANSLATE TEXTS FROM HINDI TO ENGLISH

Pranshul Pahwa

ABSTRACT

Objectives: To provide approaches for effective Hindi-to-English Machine Translation (MT) that can be helpful in inexpensive and ease implementation of and MT systems. Methods/Statistical Analysis: Structure of the Hindi and English languages have been studied thoroughly. The possible steps towards the Natural languages have also been studied. The methods, rules, approaches, tools, resources etc. related to MT have been discussed in detail. Findings: MT is an idea for automatic translation of a language. India is the country with full of diversity in culture and languages. More than 20 regional languages are spoken along with several dialects. Hindi is a widely spoken language in all the states of country. A lot of literature, poetries and valuable texts are available in Hindi which gives opportunities to retranslate into English. However, new generation is learning English rapidly and also showing keenness to learn it in simplified lucid manner. Several efforts have been made in this direction. A large number of approaches and solutions exist for MT still there is a huge scope. The paper addresses the challenges of MT and solution efforts made in this direction. This motivates researchers to implement new Hindi-to-English Machine translation systems. Application/Improvements: Efficient, inexpensive and ease translation for available Hindi literature, poetries and other valuable texts into English. Children can easily learn the culture through the poetries and literatures hence the Machine Translation of these will bring wonderful impact.

1. INTRODUCTION

India is one of the finest examples for multi-lingual and multi-social country. People from different regions speak different languages. After the analysis, it is found that the spoken languages may change after in every few kilometres (in digits of 10s). In India, Hindi is the national language which is spoken by most of the people. English is internationally accepted language which is used for communication throughout the world. The constitution of India accepts only these two languages Hindi and English as official languages. The official communication between central and state governments is also done in these two languages. The states government may have their own regional languages to carry out them work. Most of the newspapers are also published in various regional languages. There are 22 regional languages named “Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi (it is official also), Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu” speak in various regions. Hence there is dire and great demand for better Machine Translation systems to establish a better communication and exchange of information with other countries, states and central governments1,2. Machine Translation is the key research area in the field of Natural Language Processing (NLP). It is a computerized and automated idea, responsible for translating the text/documents from one language (called source language) to another language (called target language). The work in machine translation area has been going on for several decades but efficient machine translation is a still challenging task. In India, the market is largest for Machine Translation3. Figure 1 represents a block diagram for a simple Machine Translation system.

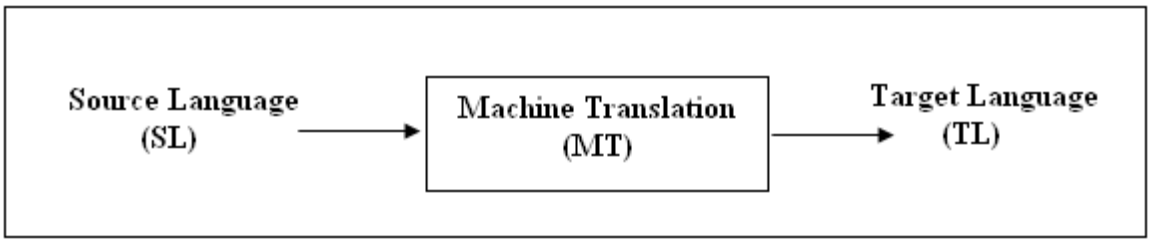

Figure 1. A simple Machine Translation (MT) System.

Machine Translation produces various challenges for all levels called "Phonetics and Phonology, Morphology, Syntax, Semantics, Pragmatics and Discourse" of Natural Language Processing. In which, ambiguity (Semantics) is the biggest one. Other than this, the different language might also have language diversity (called translation divergence) problem. Machine Translation systems deal with ambiguity and the linguistic diversity problems under the umbrella of Natural Language Processing4. In India, we feel that the important and foremost Machine Translations are HindiðEnglish and HindiðRegional Language.

1.1 Hindi-to-English Translation

Hindi is our national language. People speak different regional language but Hindi is the main official language for standard communication. Other than us, Hindi is known in other countries like Pakistan, Bangladesh and Nepal etc.

The default structure of Hindi sentence is Subject- Object-Verb (SOV), e.g.

"पृथ्वी सोना चाहता है |" where S = पृथ्वी, O = सोना and V = चाहना

Indian languages (primarily Hindi) have the following characteristics:

• Highly inflectional language,

• Rich morphology, and

• Relatively free word order.

The Hindi-to-English Machine Translation is more complex due to its characteristics. Anything written in Hindi may show different senses depending upon the context. The spoken sequence of any statement in Indian language may differ by people5,6. Figure 2 represents a block diagram for a Hindi-to-English Machine Translation system.

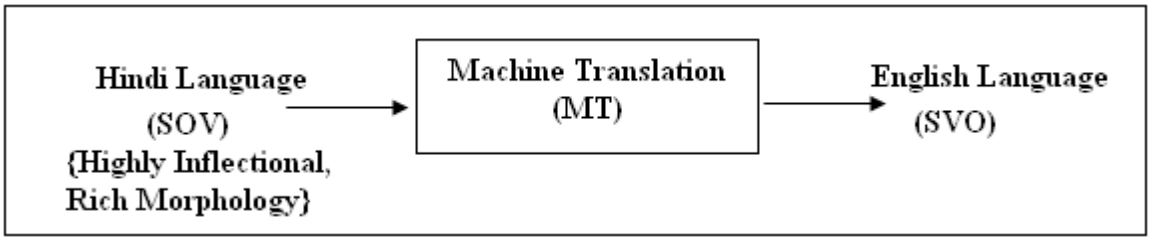

Figure 2. Hindi ð English Machine Translation.

1.2 English-to-Hindi Translation

English is a major internationally accepted language which is spoken and used in all kinds of communications among almost all countries throughout the world. We can also say that almost English is the only language which is popular among people from all over the world. The default structure of the English sentence is Subject-Verb-Object (SVO), e.g. “Prithvi wants gold” where S = Prithvi, V = want and O = gold. English is having following main characteristics:

• Highly positional language

• Rudimentary (poor) morphology.

English-to-Hindi Machine Translation results a verb movements of large distance. Hindi satisfies the gender agreement also, which is not possible in English. By enriching the source side English resources with linguistic factors, the morphological issues can be resolved5,6. Figure 3 shows a block diagram for an English-to-Hindi Machine Translation system.

The HindióEnglish Machine translation can be improved by incorporating technique called Word Sense Disambiguation. Word Sense Disambiguation (WSD) is defined as the task of identifying the correct sense of a word depending upon the context. Word sense disambiguation algorithms can be broadly classified as knowledge/ dictionary-based, supervised, semi-supervised, unsupervised approaches. However, there is no boundary in using either single or combinations. Earlier, the combinations have also produced good results7,8.

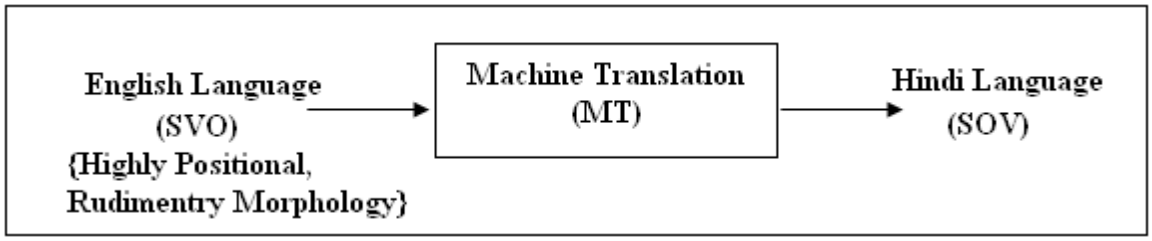

Figure 3. English ð Hindi Machine Translation.

Since last 03 decades, In India a lot of research and research projects are done in the area of Machine Translation. Although they have produced some good Machine Translation systems, they all have their own advantages, disadvantages and limitations and “It is not possible to have fully automatic, qualitative, and general purpose Machine Translation5”. Hence, still there is scope for researchers to do more research in this area. A lot of researches and research projects are also on going to overcome these disadvantages and limitations. These scopes are motivating the Teaching of Machine Translation in Indian perspective to the students and researchers9.

In the field of Machine Translation, a lot of surveys are done in the Indian perspective. First, Survey relates to resources, services and tools for Machine Translations system throughout India. This survey is the rigorous collection for the Indian perspective10. Second, Survey includes Word-sense Disambiguation approach which can be used for improving the Machine Translation system11. This contains the type of approach (like knowledge-based, supervised, minimally-supervised, unsupervised, hybrid etc.), corpus or WordNet details, features, advantages, disadvantages and limitations of the approach, new techniques under these approaches etc. Third, Survey includes different types of Machine Translation approaches used for developing the systems12-15. Surveys related to approaches include the name of approach (like direct, rule-based, corpus-based, hybrid etc.) for developing the Machine Translation system, features, advantages, disadvantages and limitations of the approach, new techniques under these approaches etc. Fourth, Survey includes different type of Machine Translation systems developed in India. Surveys related to these systems contain name, year of development, people and/or organization, funding agency, place of development,

domains/applications if the system, approaches/techniques and tools/resources used, features etc14-17. The all types of surveys also display the web-links to use these kinds of Machine Translation systems. The literature available in this paragraph is based on survey papers only but the next paragraph is based on actual research, research projects and resources.

Machine Translation system faces ambiguity and divergence issues at all levels of Natural Language Processing4,18. It is observed that the multilingual system is bounded to resource constraint like WordNet which is costly and takes more time in processing. Anglabharti is English to Indian languages machine aided translation system19. It is using rule-based (pseudo-interlingua based) method.

The system produces good results. However, sometimes produces more than one target sentences for a given source English sentence. Computer Assisted Translation System Mantra, translates the texts from English to Hindi in the domain of Personnel Administration, is developed using rule-based (transfer-based) method20. Research through this system produces new areas to contribute other facilities. The Anusaaraka system, makes documents accessible in one Indian language to another Indian language, is developed using direct (word-to-word)

method21. This system also produces good results but if it enters into common use, it has major implications. Universal Networking Language (UNL) {Interlingua}- based machine Translation system is used translation for English to Indian languages although is a good system but language divergence issues between source and target to the UNL results implications22. AnglaHindi is a participant project of the Anglabharti translation and responsible for English to Hindi translation23. It is developed using rule and example-based hybrid method. MaTra is a fully automatic system for English-Hindi Machine Translation (MT) of general-purpose texts24. It is developed using rule-based (transfer-based) method.

Statistical-based Machine Translations by Google, Microsoft, Worldlingo and IBM are Google Translate, Bing Translator, Worldlingo and IBM Server respectively. Machine Translation approaches are classified as direct translation, rule-based (transfer and Interlingua-based) translation, corpus-based (statistical and example-based) translation and hybrid (combination of one or more) translations25. These systems and approaches have them own features, advantages, disadvantages and limitations. The Statistical Machine Translation (SMT) Model3,14 and its types Word, Phrase and Hierarchical Phrase Based Models and others provides the basis to improve the Machine Translation systems. These are helpful in developing new systems also. A number of online applications are available and accessible for Hindi-to-English Machine Translation. By analysing the output, it can be easily observed that most of the applications failed to produce desired output. Only “Google Translate” is producing good result “Earth wants to sleep”. However, it cannot identify the Noun “पृथ्वी” that’s why it is producing “Earth” whether it should write “Prithvi”. The remaining applications are producing improper results. Hence, it can easily analyse that there is a need of an enhanced and appropriate version of Hindi-to-English Machine Translator which can provide better and appropriate result.

WordNet is an online lexical database designed for English language includes four main Parts-of-Speech (PoS) (i) Noun, (ii) Verb, (iii) Adjective and (iv) Adverb which are organized into sets of synonyms26. HindiWordNet is an online lexical database designed for Hindi language on the basis of English WordNet. Similar to English WordNet, It also includes the four main parts of- speech of Hindi (i) Noun, (ii) Verb, (iii) Adjective and (iv) Adverb, which are organized into sets of synonyms. IndoWordNet is a linked structure of word nets of major Indian languages27.

Word-sense Disambiguation algorithms and applications are categorized as knowledge/dictionary-based, supervised, semi-supervised, unsupervised and hybrid approaches7. They have their own features, advantages, disadvantages and limitations. The critical analysis provides the knowledge to choose the appropriate Word sense Disambiguation approach for improving the Machine Translation Systems28. Unsupervised Word Sense Disambiguation based an experimental study of Graph Connectivity helps in improving the Machine Translation29. Concept map construction might help in improving the Machine Translation

because with the help of this, the ideas and knowledge can be combined which are related to each other in some respect. This creates a semantic binding between two ideas or knowledge. With concept map, we can interlink the concepts which belong to the same domain30,31.

Chinese-Japanese Sign Language Translation proposed system provides research directions for other kind of similar translations like HindiðEnglish Sign Language Translation System32. Bi-lingual Hindi-English (Hinglish) Machine Translation plays important research direction for separate the pure component languages from a mixed set language33.

BLEU (Bilingual Evaluation Understudy) is the major and some other metrics are helpful in the automatic evaluation of Machine Translation system. There are different techniques under BLEU which play important role in evaluation the Machine Translation system6,34.

A lot of ancient literatures exist in Hindi. They are written on "Devanagari lipi (script)" which had been developed during 15th Century. Mostly books, novels, volumes etc. are in Hindi script. In modern era, there is a huge demand for English translation. Since last decades, the research has been increased35. One of the hardest kinds of machine translation is poetry translation. A lot of poetries are available in Hindi. A lot of work has been done in this move. Available system requires better mechanism for poetry translation into English36.

Many researchers, institutions and research organizations have started working on Machine Translation systems for Hindi to English translation, English to Hindi, Hindi to regional language translation and vice-versa and have succeeded in obtaining very satisfactory results. The prominent institutions and research organizations which have worked in area of Machine Translation and still working are as follows2,5,17:

• Technology Development for Indian Languages (TDIL) project by Department of Electronics and Information Technology (DeitY), Ministry of Communications and Information Technology, Government of India.

• Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Kanpur, Bombay and Delhi.

• Department of Computer and Information Sciences, University of Hyderabad (UoH),

Hyderabad.

• Language Technologies Research Centre (LTRC), International Institute of Information Technology (IIIT), Hyderabad.

• Centre for Development of Advanced Techniques (CDAC), Pune, Noida and Bangalore.

• National Centre for Software Technology (NCST) (Now CDAC), Bombay.

• Department of Computer Science and Engineering, Jadhavpur University, Kolkata.

• Machine Learning Lab, CSA, Indian Institute of Science (IISc), Banglore.

• AU-KBC Research Centre, Chennai.

• Department of Computer Science and Application, Utkal University, Utkal.

• Advanced Center for Technical Development of Punjabi Language, Literature and Culture,

Punjabi University, Patiyala.

• Computational linguistics R&D, Jawaharlal Nehru University (JNU), New Delhi, etc.

A large no of private organizations and industries have also played important role in the development of various Machine Translating systems for Indian languages by integrating these into their worldwide projects. The prominent private institutions and research industries which have been worked in area of Machine Translation and still working are as follows:

• Google - Google translator supports Machine Translation for more than 85 languages.

• Microsoft - Microsoft translator supports Machine Translation for more than 45 languages.

• Worldlingo – Worldlingo translator supports Machine Translation for more than 33 languages.

• IBM - IBM Translator is currently available for Machine Translations among English, Brazilian Portuguese, Spanish, French and Arabic languages only.

Besides, these Government and Private organizations, there are a number of other organizations who have worked in Machine Translation and still working. Some of them are the funded projects also.

2. POSSIBLE APPROACH

Ambiguities and Translation Divergences (TD) are very challenging issues for any machine translation system. Here, we are trying to propose a possible approach which may resolve these issues and provide precise and quality translation. Our approach for Hindi ó English Machine

Translation will be constituting the following seven modules and their use with functionalities shown in system architecture Figure 4. The modules are:

a. Source Language (SL): Hindi

b. Target Language (TL): English

c. Stages for Natural Language Processing and ambiguities:

• Morphology (Word forms based processing)

• Lexicon (Words Storage and their associated knowledge)

• Parsing (Structure and/or syntax based processing)

• Semantics (Meaning based processing)

• Pragmatics (Human intention and model based processing)

• Discourse (Connected text based processing) 4,7

d. Machine Translation (MT) Methods/Models:

• Direct MT

• Rule(Transfer & Interlingua)-based MT

• Corpus(Example & Statistical)-based MT

• Hybrid MT25

e. Word-Sense Disambiguation (WSD) approaches:

• Dictionary and Knowledge-based WSD

• Supervised WSD

• Semi/Minimally-Supervised WSD

• Unsupervised WSD

• Hybrid WSD7

f. Concept-Map Construction for

• Source Language (SL)

• Target Language (TL) 30,31

g. WordNets

• WordNet

• HindiWordNet

3. SYSTEM ARCHITECTURE

The system architecture of a possible approach with seven modules and their functionalities in Hindi ó English Machine Translation has been shown in Figure 4. This is proposed possible approach and in the near future, it can be further modified if needs or requirement changes.

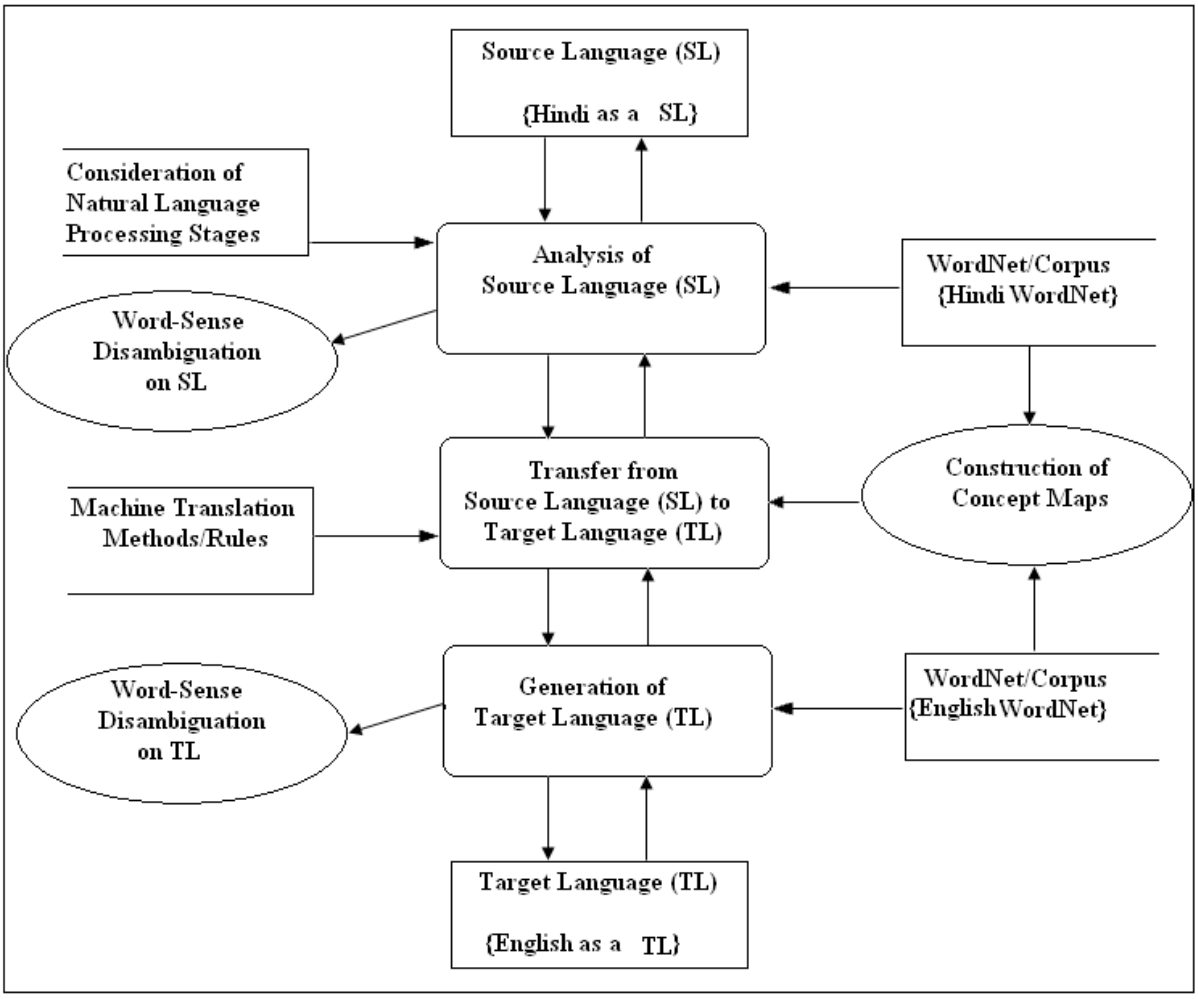

Figure 4. System Architecture for Functionality of Hindi ó English Machine Translation37.

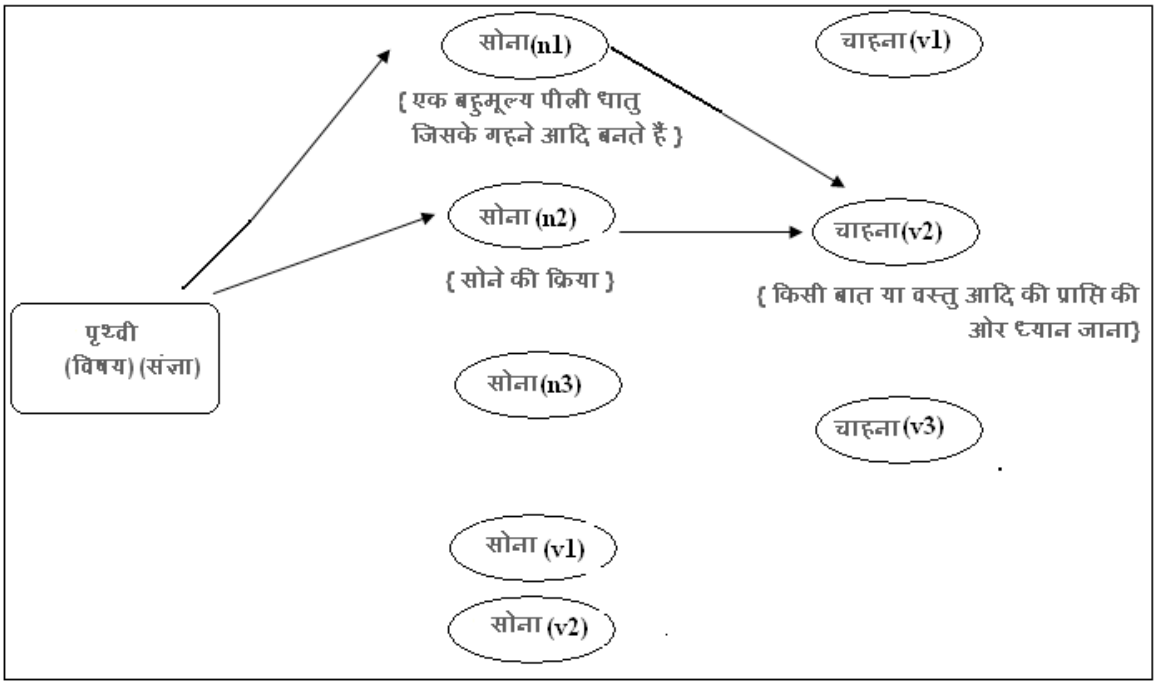

Figure 5: Word-Sense Disambiguation and Concept-Map Construction for Hindi Sentences " पृथ्वी सोना चाहता है |"29-31

The examples of use of Word-Sense Disambiguation algorithms and Construction of Concept-Maps for the appropriate HindióEnglish Machine Translations have been shown in Figure 5, 6 and 7. We have drawn the graphs on WordNet and HindiWordNet for both English and Hindi Language respectively. By analyzing these graphs, we can assume that Word-Sense Disambiguation algorithms and Concept-Map Construction will definitely helpful in Hindi ó English Machine Translations.

4. CONCLUSION

In this comprehensive survey 37 considered articles reviewed rigorously for various HindiðEnglish and other Machine Translation projects available in India and abroad. The techniques, approaches and resources for development of Machine Translation systems have been studied meticulously. Preliminarily analysis has been done by running examples on various existing Machine Translation Systems. Various models related to Machine Translation have been studied thoroughly and architecture has been proposed which might help in improving the existing Machine Translation systems. The expected outcomes of proposed approach may be as follows:

• Models for Machine Translation of Hindi poetries (and/or) literature into English.

• Machine Translated HindiðEnglish statements with higher and improved precision.

• Solution for Machine Translation of large shallow depth poetries (and/or) literature.

• High quality of Machine Translated systems for HindiðEnglish.

5. FUTURE SCOPE

In the present work, importance of Machine Translations has been discussed deeply and an approach for an effective Hindi-to-English Machine Translation (MT) has been provided that can be inexpensive and ease implementation of and Machine translation systems. For the same, various kinds of Hindi tó English Machine Translation systems have been studied and analysed throughout India as well as outside India. Multiple Resources, Techniques and Tools have been discussed which can be used in the implementation of these kinds of Hindi tó English Machine Translation systems and their enhancement. The proposed approach and architecture gives a new direction for the betterment of the Hindi tó English Machine Translation systems. In near future, implementation aspects will be covered for Hindi tð English Machine Translation System for Primary Education through poetry translation. Structure of Poetries and Literature has been studied thoroughly for the same. High quality poetry translation from Hindi tó English might be there which may consider as the best kind of Poetry Translation in the world.