CHAPTER 3

REVIEW OF LITERATURE

India is a linguistically rich nation with 22 official languages written in different scripts. The official and national language of India is Hindi. Along with Hindi, states have different languages for official works. Despite all these official languages many states and central officials use English for official work. English is prevalent in the areas like media, science and technology and commerce.. English is also an official language of Indian judiciary system. Moreover, some states work in their regional languages, for example, in Punjab, all formal letters are in Punjabi only.

India is a diverse nation where regional languages are still preferable as compared to national or international languages. This necessitates a machine translation system to break the language barrier in the Indian community. The machine translation system development process started two decades ago for Indian languages, but this is still an ongoing task.

This session of the thesis discusses various approaches used to develop machine translation system for Indian languages. Efforts have been made to collect all the possible information regarding machine translation work done for Indian languages. The following literature review presents the history of machine translation system development for Hindi, Urdu, Punjabi and other official languages of India.

3.1 ANGLABHARTI by Indian Institute of Technology, Kanpur (1991)

This machine translation system was developed for multilingual translation from English to various Indian languages. The primary focus was English-Hindi translation (Sinha et al., 1995; Manning et al., 2003). The approach used in this system was better than transfer-based approach but not as accurate as compared to interlingua approach. The system was divided into different phases to perform complete translation, for example, the first phase, was parsing of patterns in input sentence of source language English to generate pseudo-target symbols. These pseudo target symbols were according to the target Indian languages. Semantic tags were used to resolve the word sense disambiguates in the source language.

Text generator modules were used to transform target language pseudo-symbols or partial target language text to the corresponding target language. Post editing module was used to make the final changes in generated text. This system was developed as a general purpose system, which was able to translate text from various domains but it is more prevalent in health domain text translation. English to Hindi translation system known as Angla Hindi, a web-based system available at (http://anglahindi.iitk.ac.in) shows good accuracy for health related data translation. The system framework research is extended for English to Tamil and English to Telugu translation. The project is primarily based at IIT-Kanpur, in partnership with ER&DCI, Noida, and funded by TDIL. Professor RMK Sinha, Indian Institute of Technology, Kanpur was leading this MT project.

3.2 ANGLABHARTI -II by Indian Institute of Technology, Kanpur (2004)

This machine translation system is an extended form of previously developed ANGLABHARTI-I system. ANGLABHARTI-II system was drawn up to overcome the disadvantages of ANGLABHARTI-I machine translation system(Sinha et al., 2003). A generalized example-based module in addition to a raw example based one was used to improve the accuracy of the translation system. Unlike the previous system both generalized example based and raw example based modules were applied on input text for further processing. Various sub-modules along with paraphrasing and automatic pre-editing steps were used to improve the overall accuracy. The ultimate aim of this machine translation system was to cover twelve official Indian languages. Various renowned institutes of India were participated in the project.

3.3 ANUBHARATI by Indian Institute of Technology, Kanpur (1995)

ANUBHARATI machine translation system was developed for Hindi to English machine translation. The system was designed using the template for hybrid example based machine translation system which is a variation of the example-based system (Antony P.J 2013, Sinha et al., 2003). This hybrid example-based approach used advantages of patterns and example-based approaches. This machine translation was developed as a generic system which could be extended to other Indian languages by adding or changing a few modules.

3.4 ANUBHARATI-II by Indian Institute of Technology, Kanpur (2004)

ANUBHARATI-II is an improved version which incorporated many different modules (Antony P.J 2013) to upgrade the overall performances of the baseline system. The system was developed using hierarchical example based approach combined with various other modules to make it hybrid machine translation system. The system was developed as a generalized system that could be extended to many other Indian languages from Hindi.

3.5 Anusaaraka by Indian Institute of Technology, Kanpur and University of Hyderabad

Most of the Indian languages are closely related to each other and these languages share a vast amount of features, for example grammatical structure, vocabulary, and morphological features. Efforts have been made to use all these features to develop a general purpose machine translation system based on the principles of Paninian grammar(Durgesh et al., 2000; Antony P.J 2013).

The system was developed for translating five regional Indian languages Punjabi, Telugu, Kannada, Bengali and Marathi into Hindi. The system has two main modules, one was Core Anusaaraka and the second was the domain-specific module. Core Anussaraka is based on language-specific knowledge and the second part was based on statistical approach or knowledge gathered using statistical models. The main idea behind this was to divide the load of the system into two different parts; fist part carry out all the source language based analysis, and the other part works as knowledge-based analysis. The Anusaaraka project was started at IIT Kanpur and funded by TDIL. It was handed over to Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad, for further development. IIT Hyderabad later developed an English to Hindi machine translation system using the architecture of the Anusaaraka approach.

3.6 Anusaaraka System from English to Hindi

Anusaaraka machine translation system was developed based on Anusaaraka architecture and load balancing features (Manning et al., 2003; Bharati et al., 1997). The XTAG based tagger and light dependency analyzer was used to analyze source language text.

3.7 MaTra (2014)

MaTra was a general purpose machine translation system for English to different Indian languages but presently working only for Hindi. It is a Human-Assisted translation system developed using transfer-based approach (Antony P.J 2013; Manning et al., 2003). MaTra provides interactive GUI to analyze the system and presents disambiguate using rule-based and heuristics approaches. The system mainly applied in the domain of technical news phrases and annual reports. It has a text classification module at the front, which determines the nature of the news story (e.g. terrorism, political, economic, etc.) before operating on the given story.

Depending on the type of news, it uses a suitable dictionary. It needs considerable human assistance in analyzing the input. An additional part of the system breaks the given complex English sentence into simpler sentences, which were then analyzed and used to generate Hindi. The system can work in a fully automatic mode and generate rough translations for end users but is fundamentally meant for translators, editors and content providers. MaTra system is funded by TDIL and development was undertaken by Natural Language group of the Knowledge Based Computer Systems (KBCS) section at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai).

3.8 MANTRA by Centre for Development of Advanced Computing, Bangalore (1999)

MANTRA was a domain specific system especially to translate the domain of gazette notifications like government appointments and parliamentary proceeding summaries. The system was developed for English to Indian language and from Indian languages to English. Lexicalized Tree Adjoining Grammar(LTAG) was used to represent source language and target language sentence structures. The system preserves the formatting of input text during the translation process. The Mantra system was developed using general approach and has limited lexicon and grammar for a domain specific translation. The system was used for Hindi-English and Hindi-Bengali parliament proceedings translation.

3.9 UCSG-based English-Kannada MT by University of Hyderabad

The system was developed using transfer-based approach and was applied to translate domain specific documents, especially government circulars. The system worked on sentence level translation and required post editing to refine the translation. The system's processing was divided into two levels; in the first part, system analyzed and parsed the source language sentence using the Universal Clause Structure Grammar (UCSG) parser. In second part, translation rules, English-Kannada bilingual dictionary, and network based Kannada morphological generator were applied on the sentence to translate into the Kannada.

3.10 UNL-based MT between English, Hindi and Marathi by Indian Institute of Technology, Mumbai

Interlingua approach based Universal Networking Language(UNL) machine translation system was developed for English, Hindi, and Marathi languages at IIT Bombay. UNL is a comprehensive project of United Nations University. This project was started to develop machine translation system for all the major languages of the world. In this machine translation system, the source language is firstly converted into interlingua-based UNL form. This UNL form is common to all the target languages and can be converted to any target language text for translation purpose. The system used hypergraph (concept of nodes and relations as directed arcs) to represent sentence of the source language. The document knowledge was represented in three dimensions as word knowledge, conceptual knowledge, and attribute labels.

3.11 Tamil-Hindi Anusaaraka MT

The Anusasraka based machine translation system was developed for Tamil-Hindi translation. The system used core architecture of Anusasraka. The KB Chandrasekhar Research Centre of Anna University at Chennai is operating in the area of Tamil NLP (Antony P.J 2013; Manning et al., 2003). The group developed a Tamil-Hindi machine-aided translation system under the supervision of Prof. CN Krishnan, with an accuracy of the system was 75%.

3.12 English-Tamil machine Aided Translation system

This English-Hindi machine translation system was developed as a prototype by a group of NLP researchers (Manning et al., 2003; Dwivedi et al., 2010). The system was based on three main components; English morphological analyzer, Tamil language morphological generation and a mapping unit for translation.

3.13 SHIVA MT System for English to Hindi

This system was developed using example-based approach for English to Hindi machine translation. An experimental system was released for experiments, trials, and user feedback and is publicly available. This project was developed collectively by the Indian Institute of Science, Bangalore, and International Institute of Information Technology, Hyderabad, in collaboration with Carnegie Mellon University(Antony P.J 2013; Dwivedi et al., 2010).

3.14 SHAKTI MT System for English to Hindi, Marathi and Telugu

This hybrid approach based system was developed for English to Hindi, Marathi and Telugu languages. The system used Rule-based and statistical based approach to develop the hybrid architecture. An experimental system for English to Hindi, Marathi, and Telugu is publicly available for experiments, trials, and user feedback. The system was developed by the joint venture of Indian Institute of Science, Bangalore and International Institute of Information Technology, Hyderabad in collaboration with Carnegie Mellon University(Dwivedi et al., 2010).

3.15 Anuvadak English-Hindi MT

The Private sector company Super Infosoft Pvt. Ltd developed this English to Hindi machine translation system(Antony P.J 2013; Manning et al., 2003; Dwivedi et al., 2010). This is developed as a general-purpose system which can be extended to different languages. The system contained various inbuilt dictionaries related to different domains and support post-editing. The system has a transliteration sub-module to handle the unknown words if system failed to find a particular word in the dictionary. The system can run on Windows platform, and a demonstration version of the system is freely available in public domain.

3.16 English-Hindi Statistical MT

IBM India worked on English-Hindi machine translation system using statistical-based approach. The system was developed for English to various Indian languages, but the main focus was English Hindi translation(Antony P.J 2013; Manning et al., 2003).

3.17 English-Hindi MAT for news sentences

An English-Hindi machine translation system was developed by Jadavpur University, Kolkata. The system was developed using transfer-based approach(Antony P.J 2013). The system was drawn up to translate domain specific text such as news sentences.

3.18 A Hybrid MT system for English to Bengali

A hybrid machine translation system for English to Bengali was developed at Jadavpur University, Kolkata, in 2004 under the supervision of Prof. Sivaji Bandyopadhyay (Dwivedi et al., 2010). The system works at the sentence level.

3.19 Hinglish MT system

To translate standard Hindi to English text, a machine translation system was developed using ANGLABHARTI-II and ANUBHARTI-II systems(Antony P.J 2013). The system was developed under the supervision of R. Mahesh K. Sinha and Anil Thakur. The system yielded 90% accuracy in all given cases except the case with polysemous verbs due to the very shallow grammatical interpretation used in the process. The system was inadequate to resolve their meaning.

3.20 English to (Hindi, Kannada, Tamil) and Kannada to Tamil language-pair EBMT system (2006)

The machine translation system was developed by (Balajopally et.al 2006) for English to Hindi, Kannada and Tamil languages as well as Kannada to Tamil language pair. The system was developed using example-based machine translation system. Various type of dictionaries were part of the system to provide solid knowledge base like bilingual dictionaries, phrase-dictionary, word-dictionaries, a sentence-dictionary and phonetic dictionary of parallel corpora of sentences along with the phrase, words and phonetic mapping of words. The size of the corpus was 75,000 commonly used sentences.

3.21 Punjabi to Hindi Machine Translation System (2007)

Punjabi to Hindi machine translation system (G.S Josan G.S Lehal 2008) was developed by G.S Josan and G.S Lehal. In this system, direct word to word mapping scheme has been used to translate Punjabi text to Hindi text. Along with direct mapping, the system contained many different modules like word sense disambiguation and transliteration module to handle unknown words which were not part of knowledge base and post-processing method was there to refine the final output. The author claimed 92.8% accuracy for this system.

3.22 MT System among Indian language - Sampark (2009)

Various academic institutions came together and made a joint venture to develop a machine translation system called 'Sampark' for Indian languages. Institutions like IIIT Hyderabad, University of Hyderabad, CDAC (Noida, Pune), Anna University, KBC, Chennai, IIT Kharagpur, IIT Kanpur, IISc Bangalore, IIIT Allahabad, Tamil University, and Jadavpur University were part of this collaboration. A team of these institutions released an experimental system for a few languages of India like Punjabi, Urdu, Tamil, and Marathi to Hindi and Tamil-Hindi translation system in 2009 (Dwivedi et al., 2010).

3.23 English to Bengali (ANUBAAD) and English to Hindi MT System by Jadavpur University (2004)

English to Bengali translation system was developed by Jadavpur University(Antony P.J 2013). The system was developed using example-based approach, and they named it ANUBAAD translation system. The system was trained on domain-specific examples and worked at a sentence level. The University researchers also worked on Bengali and Manipuri to English machine translation. The university used these translation systems for guiding the researchers who worked in the machine translation area. These translation systems were developed under the guidance of Prof. Sivaji Bandyopadhyay.

3.24 Oriya MT System (OMTrans) by Utkal University, Vanivihar

An English-Oriya machine translation system was developed by Utkal Univeristy, Bhuvaneshwar. This English-Oriya machine translation system was named OMTrans. It was developed under the supervision of Prof. Sanghamitra Mohanty (Manning et al., 2003). The system contained N-gram based word sense disambiguation sub-modules. The system also contained a parser and an Oriya Morphological Analyzer.

3.25 English-Hindi EBMT system by IIT Delhi

This English-Hindi machine translation system was developed by the Department of Mathematics, IIT Delhi, under the supervision of Professor Niladri Chatterjee(Antony P.J 2013). The system was developed using example-based approach. They developed divergence algorithms for recognition of English to Hindi example-based system and a systematic scheme for retrieval from the English-Hindi example base.

3.26 Machine Aided Translation by Centre for Development of Advanced Computing (CDAC), Noida

This Machine Aided Translation system approach has been used to develop English to Hindi translation. The system has the capability to translate domain specific text, for example health-related data(Manning et al., 2003). The system used post editing sub-module to improve the accuracy and reported 60% accuracy on the domain-specific text.

3.27 Hindi to Punjabi MT system (2009)

The Rule-based Hindi to Punjabi Machine Translation system has been developed by (Vishal Goyal, G.S Lehal 2008). The system was developed using direct mapping or word to word mapping technique which is a basic form of rule-based approach. Along with word-to-word mapping approach the system used various rules and sub-modules like morphological analyzer and named entity recognition modules(Goyal et al., 2009; Dwivedi et al., 2010). The system used word sense disambiguation module to find the correct sense of the word before mapping to the target language word. To handle the unknown words, transliteration module was added to the system. The accuracy of the system was reported to be 95.40% by Intelligibility test and 87.60% on the basis of accuracy test.

3.28 A Statistical MT Approach to Sinhala-Tamil Language (2011)

Sinhala-Tamil machine translation system was developed using statistical machine translation system by Ruvan Weerasinghe (Antony P.J 2013). The statistical machine translation approach was also applied on Sinhala-Tamil and English-Sinhala pairs. Experiments show that statistical approach shows better results for Sinhala-Tamil language pair than the English-Sinhala language pair. The author also mentioned that statistical approach works better for languages that are not too distantly related to each other.

3.29 An Interactive Approach for English-Tamil MT System on the Web (2002)

An interactive approach has been used to English-Tamil language pair to develop the machine translation system (Antony P.J 2013). The system was developed by Dr. Vasu Renganathan, University of Pennsylvania. The system contained a set of rules and knowledge-base contained five thousand words. Various transfer based rules have been used to map English sentence forms to Tamil sentence structures. The system was developed for the web-based platform and had the interactive system to add more words in the knowledge base and rule to the rule-based module.

3.30 Translation system using pictorial knowledge representation (2010)

A latest machine translation approach was introduced by Samir Kr. Borgohain and Shivashankar B. Nair. This translation approach was based on their pictorial knowledge for Pictorially Grounded Language (PGL) (Antony P.J 2013). In this method, symbols of both the source and the target languages were grounded on a common set of pictures and animations. PGL was a graphic language and represents a traditional intermediate language illustration. While preserving the inherent meanings of the source language, the translation mechanism can also be scalable into a larger set of languages. The translation system was implemented in such a way that pictures and objects were tagged with both the source and target language equivalents, which produces the reverse translation much simpler.

3.31 Hindi to Urdu transliteration system (2010)

G. S.Lehal and T. S.Saini (2010) developed Hindi to Urdu transliteration system developed at Punjabi University, Patiala with high accuracy of 99.46% at the word level. The system tried to overcome the shortcomings of the existing rule-based Hindi to Urdu Transliteration systems (G.S Lehal and T.S Sani 2010). Various challenges such as multiple/zero character mappings, variations in pronunciations and orthography, transliteration of proper nouns, Urdu word boundary, etc. have been handled by generating special rules. Also, it used various lexical resources such as Hindi spell checker, Urdu and Hindi word frequency lists, Urdu word bigram list, Hindi-Urdu lookup table, etc. The system was developed in three different phases, pre-processing modules, core processing module and post-processing module. In pre-processing module, Hindi spell checker was used to correct spelling mistakes and normalization of the input text. In core processing part Hindi-Urdu dictionary has been used with more than 12,000 unique words. Lexicon lookup process maps Hindi and Urdu words based on various linguistic rules. In post processing part, Urdu spell checker has been used to correct partially generated text. Urdu word frequency list having 2,31,344 unique words was used for Urdu spell check. In finalization step, the system tried to merge the phrases which are over-segmented. For this process, Bigram list of Urdu words was developed with frequency count.

3.32 Rule-based Reordering and Morphological Processing For English-Malayalam SMT (2009)

English to Malayalam machine translation system was developed by M.Tech students under the guidance of Dr. K.P Soman (Rahul et al., 2009). The system has been developed using statistical-based approach. Along with statistical approach rule-based approach has been implemented for reordering and morphological analysis of source and target language to improve the overall accuracy.

3.33 SMT using Joshua (2011)

English to Telugu machine translation system was developed using statistical-based approach. The developers Anitha Nalluri and Vijayanand Kommaluri called this system 'enTel'. The system was based on Johns Hopkins University Open Source Architecture (JOSHUA) (Antony P.J 2013). To train the statistical models, Telugu parallel corpus from the Enabling Minority Language Engineering (EMILLE) developed by CIIL Mysore and English to Telugu Dictionary, developed by Charles Philip Brown, were used to develop the translation system.

3.34 Multilingual Book Reader(2006)

A multilingual book reader interface for DLI that promotes transliteration and good enough translation was developed by The NLP team, including Prashanth Balajapally, Phanindra Bandaru, Ganapathiraju, N. Balakrishnan, and Raj Reddy (Antony P.J 2013). The system was based on direct or word-to-word rule-based translation-based approach. The system has been developed to translate between various official languages of India. This is a straightforward, simple tool that exploits the similarity between Indian languages. This system can be helpful for people who can understand their mother tongue or other Indian languages but cannot read the script, and for an average reader who has the domain expertise. This system can also be used for translating either the documents or the queries in a multilingual search purpose.

3.35 A Hybrid Approach to EBMT for English to Indian Languages (2007)

The example-based machine translation system was developed for English to various Indian languages by Vamshi Ambati and U Rohini. The system used other methods along with baseline system to make the approach hybrid(Antony P.J 2013). Authors developed various dictionaries using statistical tools. These tools were applied to the parallel corpus of various other languages.

3.36 SMT by Incorporating Syntactic and Morphological Processing

A statistical approach based system for English to Hindi translation was proposed by Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah, and M. Sasikumar to improve the overall performance of translation system (Antony P.J 2013). Authors suggested morphological and syntactic approach to be incorporated in baseline statistical system to improve the accuracy of the translation. The baseline system enhanced the performance by using recording and handling suffixes of target language words.

3.37 Prototype MT System from Text-To-Indian Sign Language (ISL)

Tirthankar Dasgupta and Sandipan Dandpat proposed this machine translation system for deaf people. This system mainly targeted Indian users only because it was based on English to Indian Sign Languages translation (Dasgupta et al. 2008; Antony P.J 2013). At present, a prototype version of English to Indian Sign Language has been developed, and the ISL syntax was represented based on Lexical Functional Grammar (LFG) formalism.

3.38 An Adaptable Frame based system for Dravidian language Processing (1999)

For Dravidian languages, a different approach that makes use of the karaka relations for sentence comprehension has been used in frame-based translation system(Idicula et al., 1999; Antony P.J 2013). Two experiments have been conducted for pattern-directed application oriented, and the same meaning representation technique was used in both cases. The first experiment was carried out to handle the free word order in the source language and convert it into fix word order. In the second experiment, the target language is an artificial language with a rigid syntax. Even though there is a difference in the creation of the target sentence, the results obtained in both experiments were encouraging.

3.39 English-Telugu T2T MT and Telugu-Tamil MT System (2004)

English to Telugu and Telugu to Tamil machine transition system was developed by CALTS in collaboration with IIIT, Hyderabad; Telugu University, Hyderabad; and Osmania University, Hyderabad under the supervision of Prof. Rajeev Sangal (CALTS). A lexicon of size 4200 words were used for English-Telugu translation system and word form synthesizer for Telugu (Antony P.J 2013). Various sub-modules and resources have been used for Telugu-Tamil translation system like Telugu Morphological analyzer, Tamil generator, verb sense disambiguator, and Telugu-Tamil machine-aided translation dictionary. The system can handle complex sentences and yield good accuracy.

3.40 Developing English-Urdu MT Via Hindi (2009)

The English to Hindi machine translation system has been used to develop English to Urdu translation system by R. Mahesh K. Sinha. Urdu and Hindi are closely related languages and share a vast number of features from vocabulary to sentence structure (R. Mahesh et al., 2009). For this English to Hindi translation system, extensive database was created which contains words and phrases. These words and phrases were further improved by adding morphological variations and post-positions. If system yields good accuracy on English-Hindi language pair, then the output of this system can be converted into Urdu text by using any extra module like part of speech tagger, chunker or parser, etc.

3.41 Bengali-Assamese automatic MT system-VAASAANUBAADA (2002)

Bengali to Assamese machine translation system was developed by Kommaluri Vijayanand, S. Choudhury, and Pranab Ratna. The system was developed using example-based approach (Antony P.J 2013). They created the parallel corpus of Bengali-Assamese to train the system. The preprocessing modules were used to process the longer sentences and help to improve the overall accuracy. The grammatical structure of Bengali and Assamese is quite similar which also helps to improve the overall accuracy.

3.42 Phrase based English-Tamil Translation System by Concept Labeling using Translation Memory (2011)

A phrase-based approach was used by The Computational Engineering and Networking research center of Amrita School of Engineering, Coimbatore to develop English-Tamil translation system (Antony P.J 2013). Parallel corpus was used to developed phrase based knowledge base. The system contains large database for translation, for example, it has 50,000 parallel sentences for training and a lists of 5000 proverbs along with 1000 idioms and phrases. The system contains more than 2,00,000 technical words and 100,000 general words. The system reports 70% accuracy.

3.43 Rule-based Sentence Simplification for English to Tamil MT System (2011)

This is an improvement in English to Tamil machine translation system to handle complex sentences(Poornima et al., 2011). To handle complex English sentences, rule-based techniques were added in the baseline system. Using rules-based approach, complex sentences are broken down into sub-sentence. In this process, the meaning of original sentences remains the same. The simplification process is embedded into the system as the pre-processing module for machine translation where the initial splitting is based on delimiters and the simplification is based on connectives.

3.44 Manipuri-English Bidirectional SMT Systems (2010)

Manipuri to English and reverse machine translation system has been developed by Thoudam Doren Singh and Sivaji Bandyopadhyay (Doren Singh et al., 2010). The statistical machine translation approach was used along with morphology and dependency relations to develop this system. The system was trained on domain-specific parallel corpus of 10350 sentences from news corpus and the system was tested on 500 sentences. This set of the parallel corpus is not enough for any statistical translation system.

3.45 English to Kannada SMT System (2010)

English to Kannada machine translation system has been developed by P.J. Antony, P. Unnikrishnan, and Dr. K.P Soman. The system was based on statistical machine translation based approach (Unnikrishnan et al. 2010). A few variations were tried In order to increase the performance of the translation system. The authors proposed reordering the source language sentences according to Dravidian syntax and using root suffix separation according to Dravidian syntax along with morphological features that helped to decrease the size of training corpus. The system reported significant improvement in results by incorporating reordering and morphological futures in baseline system. Authors claimed that system worked well on simple English sentences having different forms of tenses as well as negative sentences.

3.46 Anuvadaksh

The system has been able to translate given English text into six different Indian languages, Hindi, Urdu, Oriya, Bangla, Marathi, and Tamil(Antony P.J 2013). The system was developed using hybrid approach and the system was able to work with different platforms and technology independent modules. The system was developed for domain specific text like tourism and built to facilitate multi-lingual community. In the hybrid approach, the system integrates six different technologies, Tree Adjoining Grammar based machine translation, statistical machine translation, Analyze, and Generate rules-based machine translation and example-based machine translation approach.

3.47 English to Assamese MT System(2007)

This machine translation system has been developed for English-Assamese language pair(Antony P.J 2013). The system's graphical user interface has been updated and redesigned. Various modification were made in Java modules to achieve this functionality. The baseline system was used Susha encoding scheme. In addition to this Assamese fonts were used according to Susha font set. The system was upgraded to display properly consonants, vowels, and matras of Assamese characters. Assamese keyboard mapping was added according to the Roman character set. The system was upgraded by entering Assamese words (the equivalent of English words) in the database (nouns and verbs).

The system was developed using rule-based approach and using bilingual English to Assamese dictionary. The machine translation system has dictionaries of Assamese text generated from English text. The gazetteers contained information about morphological, syntactic and partially semantic information. The system has the ability to translate simple English sentences to Assamese sentences with good accuracy but cannot handle complex sentences. The knowledge of system contains 5000 root form of words. The system translates given source language texts to the similar target language texts phrase to phrase using the bilingual dictionary lookup.

3.48 Tamil-Malayalam MT System

Tamil-Malayalam machine translation system has been developed by Bharathidasan University, Tamilnadu(Antony P.J 2013). The system consists of various modules like the lexical database, suffix database, morphological analyzer and syntactic analyzer. Lexical data contains root form of words. All forms of nouns and verbs were collected in the database. Inflectional suffixes, derivational suffixes, plural markers, tense markers, sariyai, case suffixes, relative participle markers, and verbal participle markers were collected in suffix database. Morphological Analyzer was designed to analyze the constituents of the words. This module helps to segment the words into stems and inflectional markers. The syntactic analyzer module finds the syntactic categories, like Verbal Phrase, Noun Phrase, and Participle Phrase.

3.49 English to Urdu MT system (2007)

Antony P.J 2013 was developed Rules Based English to Urdu MT system, This machine translation system was based on transfer-based approach. The system handled case phrase, verbs and posted position through Paninian grammar. Stanford parser was used to create the parse tree of given input sentences.

3.50 Hindi To Punjabi MT system(2013)

Ajit Kumar and Vishal Goyal (2013) developed Hindi to Punjabi Machine Translation system. The system was developed using phrase base statistical approach. The system was trained on a large number of parallel corpora. Most of the corpus was created manually and using various tools along with already developed machine translation system for Hindi to Punjabi.

The system was developed using various statistical tools like Giza++ for generating the alignment and phrases from the parallel corpus and MOSES toolkit which is a decoder tool to process input sentences and find proper phrases in the target language.

Along with these statistical tools, tokenization module was developed for Hindi text which is not available in MOSES toolkit. The system has the ability to handle unknown words. The Hindi to Punjabi transliteration system was developed to change the script of unknown source language words. Hindi and Punjabi are closely related languages, and transliteration plays a very effective role to handle these unknown words and increase the overall accuracy of the system.

The system was developed for the web-based and Linux based platform. Giza++ and MOSES toolkits are available for Linux platform only. Therefore, the system can be deployed on Linux based systems only, and there is no direct way to install these tools on Windows platform. The web version of the system is working on two different locations, http://tdil-dc.in/hi2pu/index.cgi and http://statmt.org/~vishal/hp/index.cgi. The first location is on the server of the Department of Technology Development for Indian Languages (TDIL), Ministry of Communication and Information Technology, Government of India. The second location is on the server of the University of Edinburgh, UK, the parent University where Moses was developed by Hieu Hoang and Philip Koehn. The authors claimed that the system is one of the most accurate translation systems for Hindi Punjabi language pair.

3.51 Shamukhi to Gurmukhi Transliteration system (2008)

T.S Saini and G.S Lehal developed Shahmukhi to Gurmukhi machine transliteration system. This system contained many different modules to handle ambiguities and refinement of output text(T.S Sani and G.S Lehal). The corpus-based approach was used to develop this system. A large corpus was used for analysis, and various statistical techniques were implemented for this propose. The corpus was collected for Shahmukhi and Gurmukhi. The statistical analysis was based on character level, word level and n-gram frequency.

Both Shahmukhi and Gurmukhi languages are two forms of Punjabi, but the main difference is the script. Where Shahmukhi uses Arabic script for writing, Punjabi use Gurmukhi script. There are many different mapping for both script characters like one-to-one, one-to-many, and many-to-one. Therefore, the system may generate ambiguous words in target script. Various rules were developed to choose perfect transliteration based on analysis of both script corpuses. The system generates all forms of transliteration and chooses the best word in target script after applying different rules and refinement techniques.

The system's knowledgebase contained many dictionaries and lexical resources, for example; Shahmukhi corpus contained 3.3 million words and Gurmukhi corpus contained 7 million words. The analysis of Gurmukhi corpus takes place in pre and post-processing phases. Shahmukhi-Gurmukhi dictionary was used in the pre-processing phase and the dictionary had most frequent 17,450 words. In the corpus analysis of Shahmukhi script, System contained around 91,060 unique unigrams. Based on the likelihood of occurrence the system incorporated around 9,000 most frequent words in this dictionary. The system reported 98.27% accuracy.

3.52 Punjabi Machine Transliteration

Punjabi machine transliteration system was developed by M.G Abbas Malik at the University of Pairs. The system was developed for Shahmukhi to Gurmukhi machine transliteration system. Both Shahmukhi and Gurmukhi are Punjabi language, but with different scripts. Shahmukhi uses Arabic script to write Punjabi text while Gurmukhi is a derivation of Landa, Shardha, and Takri, old scripts of the Indian subcontinent.

The system was developed using various linguistic rules. The system contained various sub-modules like input text parser which returns Shahmukhi tokens. PMT token converter module converts the Shahmukhi token into Gurmukhi tokens using PMT rule managers which have character mapping rules and dependency rules. The author claimed 98% to 99% accuracy on classical literature and modern literature respectively. However, there are many flaws reported by T.S Saini and G.S Lehal in their research work on this system. The system was not able to handle segmentation issues checking the bigram words like izafates and related words. The system was not able to handle words without diacritical marks. This is one of the major drawbacks of this system and yield many wrong or ambiguous words in the target script.

3.53 Urdu-Hindi Transliteration system (2012)

Urdu-Hindi Transliteration system was developed by G.S Lehal and T.S Saini. Hindi and Urdu languages are same, but the scripts are different. Hindi is written in Devanagari script, and Urdu is written in Arabic script. Urdu and Hindi speaking people can understand each other very easily because both languages have same grammatical structure and share vast amount of vocabulary. The only hurdle is they cannot read scripts of each other. The script is creating a barrier between both communities.

The authors presented a complete system to transliterate between Urdu-Hindi script. The system is able to handle various known challenges of transliteration. For example, recognition of Urdu text without diacritical marks, filling the missing script maps, multiple mapping for Urdu characters, transliterate ambiguity at the word level, word-Segmentation issues, compound words in Urdu.

The system was divided into three phases; preprocessing, processing and post processing. Every module is based on various dictionaries of both languages. In preprocessing phase, modules like normalization and to join the broken words were included. This process was dependent on a large Urdu corpus. In processing phase, three different modules were implemented; convert common Urdu word or compound words to Hindi, convert remaining Urdu Word to Hindi using various sub-modules like Urdu-Hindi mapping rules, Hindi trigram character language model, Hindi unigram word language model, Urdu/Hindi stemmer, and Segment merged Urdu words. This phase is dependent on dictionaries like Urdu-Hindi parallel corpus, Hindi character trigrams language model and Hindi word unigram language model. In the last phase, the system tried to join Hindi words, disambiguate Hindi Words using Hindi word unigrams, bigrams, and trigrams. The system reported 97.74% accuracy.

3.54 Google Translator

Google Translator is a free translation service that provides instant translations between 103 different languages. Punjabi also became the part of this list in the year 2013. Google translator is a widely used translator as compared to any other translation system. This translation service is available on different platform like Android, IOS, Windows operating systems and Web site is also available. Google is also providing translation API support for the third-party application. 500 million people are using Google translator on a daily basis using different services. This system has been developed using phrase-based statistical machine translation approach. Initially, Google translation system was developed using rule-based approach. However, since 2007 Google completely shifted to the statistically based system.

Web-based Google translator system has text speech future to read translated text. This system has an interactive interface which highlight the translated word or phrase corresponding to source language text. It can automatically detect the language of the input text. Google is also providing the facility to translate web page on the support. Brower extinction is also there for different popular browsers like Firefox and Chrome etc.

Google Translate generates a translation by looking for patterns in hundreds of millions of documents to decide best translation. By identifying patterns in documents that have already been translated by human translators, Google Translate makes guesses as to what a relevant translation should be. Currently, Google translator is based on phrase-based statistical machine translation. Urdu to Punjabi MT results are not up to the mark because Punjabi is a newly added language to this translator and the system does not have enough data for accurate translation.

Google translator is a general propose system and it does not focusing on particular language pair for translation but developed a parallel knowledge base for all the languages. The system does not directly translate text from one language to another language but is designed to translate text using one language pair to another, to get final result. for example:

Ukrainian (uk ↔ ru ↔ en ↔ other)

Urdu (ur ↔ hi ↔ en ↔ other)

Google translator has been trained on a very large corpus of documents. These documents are mainly from United Nations. Google applies various statistical based models to extract parallel phrases and words from these translator documents. United Nation documents are published in six different languages. 150 to 200 million words have been collected for translation in the bilingual pair and billions of words in monolingual collection.

Google translation system is the most popular translation system because it provides good translation for a few set of language pair. Still, Google translator is under improvement and adding many different features for translation and is upgrading itself periodically. The system has some limitations as any other translation system, it is not able to handle languages properly, which are not closely related like Urdu to English. The system makes many mistakes while doing the translation for these languages, specially in reordering and not able to handle unknown words.

Recently, Google has completed ten years, and Google claimed that it is translating 100 billion words in the day. Google translation system is updated by many people around the world. This translation community has 3.5 million members from different countries and made 900 million contributions to improving the system. Brazil is on the top of the list of countries where Google translator is most popular. Google also provides translation facility in an offline version where user do not need an internet connection.

Google is also testing artificial neural network based approach for translation known as Google Neural Machine Translation(GNMT). This approach uses state of the art training techniques to improve the translation quality. While Phrase-

Based Machine Translation (PBMT) breaks an input sentence into words and phrases to be translated largely autonomously, Neural Machine Translation (NMT) considers the entire input sentence as a unit for translation. The benefit of this approach is that it requires fewer engineering design choices than previous Phrase-Based Translation systems. When it first came out, NMT showed same accuracy with present Phrase-Based Translation systems on modest-sized public standard data sets. GNMT system produces translations that are vastly enhanced compared to the previous phrase-based production system. GNMT reduces translation errors by more than 55%-85% on several important language pairs. Test data was sampled sentences from Wikipedia and news websites. GNMT was applied successfully on English-Chinese language pair. In future Google will implement this approach over 10,000 language pairs supported by Google Translate.

3.55 Bing Translation system

Bing Translation system was developed by Microsoft. This machine translation system is based on statistical approach and is able to translate 53 different languages. This is a free service by Microsoft. Bing translation system included Urdu to Hindi translation which is developed and supported by JNU University. Punjabi is not supported by this system currently. The Microsoft Translator API(Application Programming Interface) is integrated across multiple consumers, developers, and enterprise products and many of them were developed by Microsoft including Bing, Microsoft Office, SharePoint, Microsoft Lync, Yammer, Visual Studio, Skype Translator, Internet Explorer, and Microsoft Translator apps for Windows Phone. Microsoft translation service is of two type paid and free. In free service, one can translate two million characters per month but in paid version billion of characters are supported for translation.

The first version of Microsoft translation system was developed between 1999 and 2000 by Microsoft Research. That system was based on semantic predicate-argument structures known as logical forms, and grammar correction rules. This system was finally used to translate the entire Microsoft Knowledge Base into Spanish, French, German, and Japanese.

Microsoft launched the web version of the translation system in 2007 and provided free service of text translation. In 2011, the service was extended to include various Microsoft Translator products through a cloud-based API, which supports products accessible to both consumer and enterprise users. An additional speech translation ability was introduced in March 2016.

3.56 Yandex Translate

Yandex translation system has been developed by a Russian IT company. Yandex was based on statistical machine translation approach. This system is available on the web site and also provides web services for standalone applications. The System currently supports 76 languages including many Indian languages like Hindi, Punjabi and Urdu. The system is also available as the app for different mobile platforms like Android, IOS, and Windows. Text to speech feature is also there; one can instruct the system to read translated text. The company is upgrading the system periodically to improve the accuracy of different language pairs. Translation of Urdu to Punjabi is not up to the mark.

Table 3.1: Machine Translation Systems for Indian Languages (Antony P.J (2013))

SNO	TRANSLATION SYSTEM	FOR LANGUAGE	Developed By	Approach	Domain
1.	ANGLABHARTI (1991)	English to Indian languages (primarily Hindi)	IIT, Kanpur	Pseudo-interli Ngua	General
2.	ANGLABHARTI - II (2004)	English to Indian Languages	IIT, Kanpur	Pseudo-interli Ngua	General
3.	ANUBHARATI (1995)	Hindi to English	IIT, Kanpur	GEBMT	General
4.	ANUBHARATI-II (2004)	Hindi to any other Indian languages	IIT, Kanpur	GEBMT	General

SNO	TRANSLATION SYSTEM	FOR LANGUAGE	Developed By	Approach	Domain
5.	Anusaaraka (1995)	Punjabi, Bengali, Telugu, Kannada, and Marathi to Hindi.	IIT, Kanpur and University of Hyderabad	PG	General
6.	Anusaaraka (1995)	from English to Hindi	IIT, Kanpur and University of Hyderabad	PG	General
7.	MaTra (2004)	English to Indian languages (at present Hindi)	CDAC, Mumbai	Transfer Based	General
8.	MANTRA (1999)	English to Indian languages and Reverse	English to Indian languages and Reverse	TAG	Administration, office orders
9.	UCSG-based MT	English-Kannada	University of Hyderabad	transfer based	government circulars
10.	UNL-based (2003)	Between English, Hindi, and Marathi	IIT, Mumbai	Interlingua	General
11.	Tamil-Hindi Anusaaraka MT	Tamil-Hindi	KBC Research Centre, Anna University,	PG	General
12.	English-Tamil HAMT	English-Tamil	NLP group	HAMT	General
13.	SHIVA (2004)	English to Hindi	IISc- Bangalore, IIIT Hyderabad, and Carnegie Mellon University	EBMT	General
14.	SHAKTI (2004)	English to Hindi, Marathi and Telugu	IISc- Bangalore, IIIT Hyderabad, and Carnegie Mellon University	RBM	General
15.	Anuvaadak	English-Hindi	Super Infosoft Pvt Ltd., Delhi	Not-Available	Not-Availa Ble
16.	English-Hindi Statistical MT	English to Indian Languages	IBM India Research Lab, New Delhi	EBMT & SMT	Not-Availa Ble

SNO	TRANSLATION SYSTEM	FOR LANGUAGE	Developed By	Approach	Domain
17.	English-Hindi MAT	English to Hindi	Jadavpur University, Kolkata	transfer based	newssentences
18.	Hybrid MT system	English to Bengali	Jadavpur University Kolkata	Hybrid	Sentence Level
19.	Hinglish MT system (2004)	Hindi - English	IIT-Kanpur	Pseudo Interlingua	General
20.	English to Indian and Kannada to Tamil language-pair EBMT system (2006)	i) English to Hindi, Kannada, and Tamil ii) Kannada to Tamil	Balajapally	Example-baseD	Most Commonly used sentences
21.	Punjabi to Hindi MT system (2007)	Punjabi to Hindi	Punjabi University, Patiala	Direct word to word	General
22.	Sampark (2007)9	Among Indian Languages	Consortiums of Institutions	CPG	Not-AvailaBle
23.	ANUBAAD (2004)	English to Bengali and English to Hindi	Jadavpur University	RBMT & SMT	News Sentences
24.	OMTrans	English-Oriya	Utkal University, Bhuvaneshwar		School book Sentences
25.	English-Hindi EBMT System	English-Hindi	IIT Delhi	Example-based, Divergence algorithms	
26.	Machine Aided Translation	English to Hindi	CDAC, Noida	Machine Aided Translation	Public health related sentences
27.	Hindi to Punjabi MT system (2009)	Hindi to Punjabi	Punjabi University, Patiala	direct word-to-word	General
28.	Sinhala-Tamil MT (2011)	Sinhala to Tamil	RuvanWeerasinghe	SMT based	General

SNO	TRANSLATION SYSTEM	FOR LANGUAGE	Developed By	Approach	Domain
29.	English-Tamil MT on Web (2002)	English to Tamil	University of Pennsylvania	rule-based	General
30.	Pictorial knowledge Based MT (2010)	English to Assamese	Samir Kr. Borgohain and Shivashankar B. Nair	pictorial knowledge	People not well versed in each other‟s languages
31.	English-Malayalam Statistical MT (2009)	English to Malayalam	AMRITA University, Coimbatore	SMT based	General
32.	enTel (2011)	English to Telugu	AnithaNalluri and VijayanandKommaluri	SMT based	
33.	Multilingual book reader interface for DLI	Translation for Indian Languages	PrashanthBalajapally and Team	Word-to-Word Translation	documents or the queries
34.	English to Indian Languages MT (2007)	English to Indian Languages	VamshiAmbati and Rohini U proposed	Example-baseD	
35.	Incorporating Syntactic and Morphological based MT	English-Hindi	Ananthakrishnan Ramanathan and Team	Stastical phrase-based	
36.	Text-To-Indian Sign Language (ISL) MT	English to Indian Language	TirthankarDasgupta, SandipanDandpat, and AnupamBau	Lexical Functional Grammar (LFG) Formalism	Deaf people in India
37.	Dravidian language Processing System (199)	Dravidian language	SumamUMAM MARY IDICULA	Adaptable Frame based	Not-AvailaBle
38.	English-Telugu T2T MT and Telugu-Tamil MT (2004)	English-Teluguand Telugu-Tamil	CALTS; IIITHyderabad; Telugu University-Hyderabad, Osmania University-Hyderabad,	Not-Available	
39.	English-Urdu MT via Hindi (2009)	English-Urdu	R. Mahesh K. Sinha	Not-Available	

SNO	TRANSLATION SYSTEM	FOR LANGUAGE	Developed By	Approach	Domain
40.	VAASAANUBAADA (2002)	Bengali-Assamese	KommaluriVijayan and S Choudhury and PranabRatna	EBMT	News
41.	Phrase based English - Tamil MT (2011)	English - Tamil	CEN, AMRITA University, Coimbatore	Phrase based	General
42.	Sentence Simplification System for English to Tamil (2011)	English - Tamil		Rule-based	
43.	Manipuri-English Bidirectional MT (2010)	Manipuri-English and -English-Manipuri	ThoudamDoren Singh and SivajiBandyopadhyay	Statistical	News
44.	English to Dravidian Language MT (2010)	English to Malayalam	CEN, AMRITA University, Coimbatore	SMT Based	Simple Sentences
45.	Anuvadaksh	English to six other Indian languages *i.e*. Hindi, Urdu, Oriya, Bangla, Marathi, Tamil	EILMT consortium	hybrid approach	Tourism
46.	Google Translate	Translations between 57 different languages	Google	SMT	General
47.	English to Assamese MT	English to Assamese		Rule-based	
48.	Russian-Tamil MT (1983-1984)	Russian-Tamil	Tamil University, Tanjore		Scientific text
49.	Tamil - Malayalam MT	Tamil - Malayalam	Bharathidasan University, Tamil Nadu	Not-Available	Not-Available

SUMMARY

From this review of various published papers and commercial systems, one can observe that Urdu and Punjabi translation system related research is very limited. There is no particularly focused research work related to Urdu to Punjabi translation system and vise versa. General purpose commercial translation systems has been developed using statistical based approach includes Urdu and Punjabi languages. Google has been working on Urdu Punjabi language pair for a few years but the system is in the development stage, and results are not up to the mark. Google translation system is a general purpose system where one approach has been implemented for different language pairs by ignoring features of language which can be help to increase overall accuracy. Yandex translation system is another system for Urdu to Punjabi Translation which is also a general purpose system, and translation quality is not good enough. Along with translation systems, a few transliteration systems are there for Urdu and Punjabi scripts. That kind of system helps to convert text to target script but cannot meet the purpose of translation.

Various approaches have been applied for Indian language pairs like direct translation, direct with the simple rule-based, transfer based, interlingua, statistical and hybrid approaches. Direct translation approach is effective only if language pair is very closely related or one can say dialect of each other. Then this approach may yield good results. Rule based approach like transfer based and interlingua cannot be used directly on Indian languages especially if languages are resource poor and these approaches are dependent upon various other modules part of speech tagger, morphological analyzer and semantic analyzer, syntactic analyzer, grammar checker and many more. This approach makes the system very complex, and languages like Urdu and Punjabi which are resource poor languages do not have resources in significant amount to develop all these sub-modules. The statistical machine translation approach is quite popular in machine translation community in these days. All the commercial translation systems have been using this approach for the last few years and get good accuracy for a few language pairs. Again, these statistical machine translation approaches can only be useful if one has a large number of parallel corpora to train statistical models. The proposed system was developed using hybrid machine translation approach, where system contains the statistical and rules-based module to refine the final translation. The hybrid approach is suitable for such resource-poor language pair, and the system can use good features of both approaches.