Literature Review

Following are the literature review presented here for the development of Hindi to Marathi Machine Translation system.

Bandyopadhyay S. (2000)

ANUBAAD translates news headlines from English to Bengali, which uses example based Machine Translation system. The input headline is initially searched in the direct example base for an exact match. If a match is obtained, the Bengali headline from the example base is produced as output. If there is no match, the headline is tagged and the tagged headline is searched in the Generalized Tagged Example base. If a match is obtained, the output Bengali headline is to be generated after appropriate synthesis. If a match is not found, the Phrasal example base will be used to generate the target translation. If the headline still cannot be translated, the heuristic translation strategy applied is - translation of the individual words or terms in their order of appearance in the input headline will generate the translation of the input headline. Appropriate dictionaries have been consulted for translation of the news headline.

R. M. K. Sinha, Jain R., Jain A. (2001)

ANGLABHARTI is a machine aided translated system designed for translating English to Indian languages. Instead of designing translators for English to each Indian language, it uses pseudo-interlingua approach. Analysis of English language only once, it creates intermediate structure – PLIL (Pseudo Lingua for Indian Languages). The PLIL structure is then converted to each Indian language through a process of text-generation. The effort for PLIL generation is 70% and text generation is 30%. Only with an additional 30% effort, new English to Indian language

translator can be built. The attempt has been made to 90% translation task to be done by machine and 10% left to the human post-editing. The project has been applied mailnly in the domain of public health.

Dave S., Parikh J., Bhattacharyya P. (2001)

UNL-based English-Hindi MT System translates using Universal Networking Language (UNL) as the Interlingua. The UNL is an international project of the Unite Nations University, with an aim to create an Interlingua for all major human languages. IIT Mumbai is the Indian participant in UNL. English-Hindi, Hindi-UNL, UNL-Hindi, English-Marathi and English-Bengali were also developed using UNL formalism.

Murthy K. (2002)

MAT is a machine assisted translation system for translating English texts into Kannada, which uses morphological analyzer/generator for Kannada. The input sentence is parsed by Universal Clause Structure Grammar (UCSG) parser and outputs the number, type and inter-relationships amongst various clauses in the sentence and the word groups. For each word, a suitable target language equivalent is obtained from the bilingual dictionary. Finally, the target language sentence is generated by placing the clauses and the word groups in appropriate linear order, according to the constraints of the target language grammar. Post editing tool is provided for editing the translated text. MAT System 1.0 had shown about 40-60% of fully automatic accurate translations. It works in the domain of government circulars.

Gore L., Patil N. (2002)

An English–Hindi Translation System based on transfer based translation approach, which uses different grammatical rules and a bilingual dictionary for translation. The translation module consists of Pre-processing, English tree generator, post-processing of English tree, generation of Hindi tree, Post-processing of Hindi tree and generating output. It works in the domain of weather narration.

Vijayanand K., Choudhury S. I., Ratna P. (2002)

VAASAANUBAADA is an Automatic Machine Translation of Bilingual Bengali-Assamese News Texts using EBMT technique. It involves Machine Translation of bilingual texts at sentence level. It includes preprocessing and post-processing tasks. The bilingual corpus has been constructed and aligned manually by feeding the real examples using pseudo code. Longer sentences are fragmented at punctuations to get high quality translation. Backtracking is used when the exact match is not found at the sentence/fragment level, leading to further fragmentation of the sentence.

R.M.K. Sinha, Jain A. (2003)

AnglaHindi is a derivative of AnglaBharti MT System for English to Indian languages, which is a pseudo interlingual rule-based English to Hindi Machine-Aided Translation System. AnglaHindi besides using all the modules of AnglaBharti, also makes use of an abstracted example-base for translating frequently encountered noun phrases and verb phrasals. The accuracy of the translation is 90%.

Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma, R. Sangal (2003)

Shakti system translates English to any Indian languages with simple system architecture. It combines linguistic rule based approach with statistical approach. The system consists of 69 different modules, out of which 9 modules are used for analyzing the source language (English), 24 modules are used for performing bilingual tasks, and the remaining modules are used for generating target language.

Bandyopadhyay S. (2004)

English-Telugu Machine Translation System uses English-Telugu lexicon consisting of 42,000 words. A word form synthesizer for Telugu is developed and incorporated in the system, which handles various complex English sentences.

R.M.K. Sinha (2004)

ANGLABHARTI-II uses a generalized example-base (GEB) for hybridization besides a raw example-base (REB). During the development phase, when it is found that the modification in the rule-base is difficult and may result in unpredictable results, the example-base is grown interactively by augmenting it. At the time of actual usage, the system first attempts a match in REB and GEB before invoking the rule-base. In AnglaBharti-II, provisions were made for automated pre-editing & paraphrasing, generalized & conditional multi-word expressions, recognition of named-entities. It incorporated an error-analysis module and statistical language-model for automated post-editing. The purpose of automatic pre-editing module is to transform/paraphrase the input sentence to a form which is more easily translatable. Automated pre-editing may even fragment an input sentence if the fragments are easily translatable and positioned in the final translation. Such fragmentation may be triggered by in case of a failure of

translation by the 'failure analysis' module. The failure analysis consists of heuristics on speculating what might have gone wrong. The entire system is pipelined with various sub-modules.

R.M.K. Sinha (2004)

ANUBHARTI-II uses hybrid Example-based Machine Translation approach which is a combination of example-based approach and traditional rule-based approach. The example based approaches emulate human-learning process for storing knowledge from past experiences to use it in future. The input Hindi sentence is converted into a standardization form to take care of word-order variations. The standardized Hindi sentences are matched with a top level standardized example-base. In case no match is found then a shallow chunker is used to fragment the input sentence into units that are then matched with a hierarchical example-base. The translated chunks are positioned by matching with sentence level example base.

Bandyopadhyay S. (2004)

Telugu-Tamil Machine Translation System uses the Telugu Morphological analyzer and Tamil generator. The system uses Telugu-Tamil dictionary developed as aprt of MAT Lexica. Also it uses verb sense disambiguator based on verbs argument structure.

Mohanty S., Balabantaray R. C. (2004)

OMTrans system translates text from English to Oriya based on grammar & semantics of the language. Architecture has various parts such as Morphological parser, POS tagger, Translator,

Disambiguator, etc. and some software tools used to see the rsult. Word Sense Disambiguation is being taken care of in this system. OMTrans uses principles of object-oriented approach.

Ananthakrishnan R, Kavitha M, Hegde J. J., Chandra Shekhar, Ritesh Shah, Sawani Bade, Sasikumar M. (2006)

The **MaTra system** used transfer approach using a frame-like structured representation. The system uses rule-bases and heuristics to resolve ambiguities. It has a text categorization component at the front determines the type of news story (political, terrorism, economic, etc.) before operating on the given story. Depending on the type of news, it uses an appropriate dictionary. It requires considerable human assistance in analyzing the input. Another novel component is that given a complex English sentence, it breaks it up into simpler sentences, which are then analyzed and used to generate Hindi. It works in the domain of news, annual reports and technical phrases.

Balajapally P., P. Pydimarri, M. Ganapathiraju, N. Balakrishnan, R. Reedy (2006)

English to {Hindi, Kannada, Tamil} and Kannada to Tamil Language-Pair Example Based Machine Translation is based on a bilingual dictionary comprising of sentence-dictionary, phrases-dictionary, words-dictionary and phonetic-dictionary. Each of the above dictionaries contains parallel corpora of sentences, phrases and words, and phonetic mappings of words in their respective files. EBMT has a set of 75000 most commonly spoken sentences that are originally available in English. These sentences have been manually translated into three of the target Indian languages, namely Hindi, Kannada and Tamil.

Google Translate (2007)

Google Translate applies the statistical machine translation approach for English to other Languages and vice-versa. Before 2007, it was using SYSTRAN for translation. The accuracy is good enough to understand the translated text, but not perfect. Google Translate is a free translation service that provides instant translations between 64 different languages. It can translate words, sentences and web pages between any combinations of supported languages.

G. S. Josan, G. S. Lehal (2008)

Punjabi to Hindi Machine Translation System is based on direct word-to-word translation approach. This system consists of modules like pre-processing, word-to-word translation using Punjabi-Hindi lexicon, morphological analysis, word sense disambiguation, transliteration and post processing. Accuracy of the system found is 90.67%. WER is 2.34% and SER is 24.26%.

Sobha L, Pralayankar P, Kavitha V. (2009)

Tamil-Hindi Machine-Aided Translation system based on Anusaaraka Machine Translation System architecture developed by Prof. C. N. Krishnan. It uses a lexical level translation and has 80-85% coverage. Stand-alone, API, and Web-based on-line versions have been developed. Tamil morphological analyser and Tamil-Hindi bilingual dictionary (~36k) are the by products of this system. It includes exhaustive syntactical analysis. They also developed a prototype of English-Tamil MAT system. Currently, it has limited vocabulary (100-150) and small set of Transfer rules.

Sampark Project (2009)

Sampark: Machine Translation System among Indian languages: Sampark uses Computational Paninian Grammar (CPG) approach for analyzing language and combines it with machine learning. Thus it uses both traditional rules-based and dictionary-based algorithms with statistical machine learning. This project has developed language technology for 9 Indian languages resulting in Machine Translation for 18 language pairs.

Chatterji S., Roy D., Sarkar S., Basu A. (2009)

Bengali to Hindi Machine Translation System is a hybrid Machine Translation system, uses multi-engine Machine Translation approach. After evaluation, BLUE score obtained is 0.2318.

Goyal V., Lehal G. S. (2010)

The **Web based Hindi-to-Punjabi Machine Translation System** is the extended version of Hindi-to-Punjabi MTS to Web. It has the several facilities including website translation, email translation, etc.

Goyal V., Lehal G. S. (2011)

Hindi-to-Punjabi Machine Translation System developed using direct word to word translation approach at Punjabi University, Patiala. The accuracy of system is 95.40% on the basis of Intelligibility test and 87.60% on the basis of accuracy test. In the quantitative tests the Word Error Rate is 4.58% whereas Sentence Error Rate is 28.82% and BLUE score found is 0.7801.

Sato S. (2009)

Web-Based Transliteration of Person Names is the web-based transliteration system of person names; from a person name written in English (Latin script), the system produces its Japanese (Katakana) transliteration extracted from the Web. Experiments have shown that the performance is sufficiently high: for 89.4% of English person names, the system produced one or more acceptable Japanese transliterations; 98.5% of system's outputs were acceptable transliterations. The system also works in reverse direction. This system was used for automatic compilation of an English-Japanese person-name lexicon with 406K entries.

Goyal V., Lehal G. S. (2009)

Hindi-Punjabi Machine Transliteration System implemented with the help of fifty-seven complex rules for making thetransliteration between Hindi-Punjabi language pair accurate after studying both the languages in details. Then rigorous testing was done by test data covering number of domains like medicine, proper names, city names, country names, colors, news related, castes, surnames, rivers, subject related technical terms etc. The system found to give accuracy of about 98%.

Vijaya, V.P., Shivapratap, K.P. CEN (2009)

Vijaya, VP, Shivapratap and KP CEN has developed **English to Tamil Transliteration system** and named it WEKA. It is a Rule based system and is used the j48 decision tree classifier of WEKA for classification purposes. The transliteration process consisted of four phases: Preprocessing phase, feature extraction, training and transliteration phase .The accuracy of this system has been tested with 1000 English names that were out of corpus. The transliteration

model produced an exact transliteration in Tamil from English words with an accuracy of 84.82%.

Ali Ijaz (2009)

English to Urdu Transliteration System based on the mapping rules. The whole process has three steps. In the first step, the mapping rules that have been used to generate Urdu text from English transcription. English text is converted to Urdu using both English pronunciation and mapping rules. In Second step, Urdu syllabification has been applied on English transcription. Consonant and Vowels have been combined to make syllable and breaking up a word into syllables is known as syllabification. To improve system's accuracy, they have applied the Urduization Rules in third step. Overall system's accuracy is 95.92%.

Sasidharan S., Loganathan R, Soman K P (2009)

English to Malayalam Transliteration system developed using Sequence Labeling Approach. The source string is segmented into transliteration units and related with the target language units. Thus transliteration problem can be viewed as a sequence labeling problem. Here the classification is done using Support Vector Machine (SVM). The model produced the Malayalam transliteration of English words with an accuracy of 90%. The corpus is includes 20,000 names for training and 1,000 names for testing.

Lehal G. S. (2009)

A **Gurmukhi to Shahmukhi Transliteration System** transliterates Gurmukhi words to Shahmukhi of Punjabi language. This system uses special rules to correct and modify spelling of

some words according to their pronounciation, lexical resources, spell checker, Gurmukhi-Shahmukhi dictionary, and Shahmukhi corpus. This system fails when words with typical spellings, which are not present in Gurmukhi-Shahmukhi dictionary and Shahmukhi corpus and Gurmkhi words with multiple spelling in Shahmukhi. It gives fairely good result 98.6%.

Malik M G Abbas, Laurent Besacier, Christian Boitet, Bhattacharyya P. (2009)

Urdu Hindi Transliteration System uses a novel hybrid approach for Urdu to Hindi transliteration that combines finite-state machine (FSM) based techniques with statistical word language model based approach. The output from the FSM is filtered with the word language model to produce the correct Hindi output. The main problem handled is the case of omission of diacritical marks from the input Urdu text. System produces correct Hindi output even when the crucial information in the form of diacritic marks is absent. The approach improves the accuracy of the transducer-only approach from 50.7% to 79.1%. The results reported show that performance can be improved using a word language model to disambiguate the output produced by the transducer-only approach, especially when diacritic marks are not present in the Urdu input.

Lehal G. S., Saini T. S. (2010)

Hindi to Urdu transliteration system developed at Punjabi University, Patiala with high accuracy of 99.46% at word level. System tries to overcome the shortcomings of the existing rule based Hindi to Urdu Transliteration systems. The 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 and using various

lexical resources such as Hindi spell checker, Urdu and Hindi word frequency lists, Urdu word bigram list, Hindi-Urdu lookup table, etc.

Josan G. S., Lehal G. S. (2010)

A **Punjabi to Hindi Machine Transliteration System** presents a novel approach to improve Punjabi to Hindi transliteration by combining a basic character to character mapping approach with rule based and Soundex based enhancements. Experimental results show that the approach effectively improves the word accuracy rate 92.65% and average Levenshtein distance 0.10 of the various categories by a large margin.

Chinnakotla M.K, Damani OM P, Satoskar Avijit (2010)

Authors have developed rule based systems for **Hindi to English, English to Hindi, and Persian to English transliteration.** They used CSM (Character Sequence Modeling) on the source side for word origin identification, a manually generated non-probabilistic character mapping rule base for generating transliteration candidates, and then again used the CSM on the target side for ranking the generated candidates. The overall efficiency by using CRF (Conditional Random Field) approach of English to Hindi is 67.0%, Hindi to English is 70.7% and Persian to English is 48.0%.

Kamal Deep, Goyal V. (2011)

The system developed by Kamal Deep & Dr. Vishal Goyal named **Punjabi to English Transliteration System** using a rule based approach and achieved accuracy of 93.23%. Transliteration scheme uses grapheme based method to model the transliteration problem. This

system addresses the problem of forward transliteration of person names from Punjabi to English by set of character mapping rules. This system is accurate for the Punjabi words but not for the foreign words. System evaluated for names from the different domains like Person names, City names, State names, River names, etc.

Kamal Deep, Goyal V. (2011)

The system developed by Kamal Deep & Dr. Vishal Goyal named **Punjabi to English Transliteration System** using Hybrid (statistical + rule) based approach preserving the phonetic structure of words as closely as possible. System transliterates person names; from a person name written in Punjabi (Gurumukhi Script), the system produces its English (Roman Script) transliteration. Experiments have shown that the performance is sufficiently high. The overall accuracy of system comes out to be 95.23%.

Josan G. S., Jagroop Kaur (2011)

Punjabi to Hindi Statistical Machine Transliteration System handles out of vocabulary (OOV) words with statistical approach. This system presents empirical results for statistical Punjabi to Hindi transliteration system. Experimental results show that statistical approach effectively improves the Transliteration accuracy rate and average Levenshtein distance of the various categories by a large margin. Normalization of spellings at the source may improve the results. The system itself could be improved by e.g. defining a better syllable similarity score, performing tuning of language model on various parameters like alignment heuristics, maximum phrase length, etc.

Reddy M. V., Hanumanthappa M. (2011)

For **English to Kannada/Telugu Name Transliteration System**, approach is based on query Translation using bilingual dictionaries with 85-96% accuracy. A new statistical modeling approach to the machine transliteration problem has presented. The parameters of the model are automatically learned from a bilingual proper name list. The model is applicable to the extraction of proper names and transliterations. This method can be easily extended to other language pairs that have different sound systems without the assistance of pronunciation dictionaries.

Jasleen Kaur, Josan G. S. (2011)

English to Punjabi Machine Transliteration System developed using statistical approach with MOSES for handling proper nouns and technical terms. Statistical Approach to transliteration is used for transliteration from English to Punjabi using MOSES, a statistical machine translation tool. After applying transliteration rules average %age accuracy and BLEU score of this transliteration system comes out to be 63.31% and 0.4502 respectively. One of major weakness of transliteration from English to Punjabi is dealing with multiple mapped Characters. One of major weakness of transliteration from English to Punjabi is dealing with multiple mapped Characters.

CDAC, Mumbai (2012)

XLIT is a transliteration tool developed by CDAC, Mumbai to convert words from English to Indian languages and back, without losing the phonetic characteristics. It can be used in Machine Translation systems, e-governance applications and other applications that need to enter text in any Indian language and English. It can be used as an input method on Linux systems as well.

XLIT adopts a statistical approach to tackle the transliteration problem. XLIT tool can be easily integrated into any desktop or web application.

CDAC, Mumbai (2012)

Rupantar is developed by CDAC, Mumbai to write in Indian languages using Roman Script. It also allows converting text from one script to other script, as 'रमेश' in Hindi to 'ரமேஷ்' in Tamil. It uses a key map based technique for writing and conversion. This tool easily integrates with other desktop and web applications. It is fast and lightweight application developed.