Sanskrit to Hindi Statistical Machine Translation System

Thesis submitted in partial fulfilment of the requirements for the award of the degree of

Master of Engineering
in
Software Engineering

Submitted By
Manmeet Kaur
(Roll No: 801631011)

Under the supervision of

Dr. Inderveer Chana
Professor

Dr. Ravinder Kumar
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA, PUNJAB, INDIA

June 2018

Certificate

I hereby certify that the research work which is being presented in the thesis entitled, "*Sanskrit to Hindi statistical machine translation system* ", in partial fulfilment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Inderveer Chana* and *Dr. Ravinder Kumar* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

(Manmeet Kaur)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Inderveer Chana)
Professor
CSED

(Dr. Ravinder Kumar)
Assistant Professor
CSED

i

Acknowledgement

This research work would be incomplete without acknowledging the people who supported and guided me for the successful completion of this work.

First of all I wish to acknowledge the benevolence of God who gave me courage and strength to face the challenges and to overcome the obstacles that occurred while working on this task.

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr. Inderveer Chana,** Professor and **Dr. Ravinder Kumar,** Assistant Professor, Computer Science & Engineering Department, Thapar Institute of Engineering and Technology, Patiala for their valuable guidance and continual encouragement throughout this research work. The appreciation and continual support they have imparted has been a great motivation to me in reaching a higher goal. Their guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

I am also heartily thankful to **Dr. Maninder Singh**, Hon'ble Head of Computer Science & Engineering Department, Thapar Institute of Engineering and Technology, Patiala for his kind support and providing basic infrastructure and healthy research environment.

I would like to express my sincere thanks to CSIR (council of scientific and industrial Research), Government of India to carry out my research work under the project titled "Sanskrit to Hindi machine translation".

I would also thank the Institution, all the faculty and staff members of Computer Science & Engineering Department, Thapar Institute of Engineering and Technology, Patiala for their direct-indirect help, cooperation and suggestions towards this work.

Last but not the least, I would like to thank my family for their wonderful support and encouragement without which none of this would have been possible.

Manmeet Kaur

(801631011)

Abstract

Machine Translation is a name given to computerized strategies used for translations of all or part of data from one regular language into another along with or without from human aid. Machine translation in the field of Natural Languages is benefiting over years of semiautomatic and manual analysis by computer programs and linguists. This is proven to be fruitful in form of publically available dictionaries and transition systems. Computer software's are used to translate data or speech from one language form to another to as to bridge the communication barrier. There are many approaches for building machine translation. The main problems in the area of Machine translation are corpus availability, training the system and searching for getting the accurate translated text. Machine translation system faces many challenges as of semantics, Structural and lexical ambiguity, word sense which needs to be taken care while building any Translation System.

In this research work, Sanskrit to Hindi statistical Machine Translation system is build using Thot, an open source tool available for Statistical Machine translation building. Thot helps in training phrase based models. The tool has many commands to follow for building of the system starting from corpus preprocessing, language and translation model training, generating of configuration files, tuning parameters, searching and end by post processing the output. The proposed system uses n-gram model for language model, phrase alignment model for translation model and branch and bond search algorithm. The system is deployed on cloud using Virtual Private Server. A parallel corpus of 1000 sentences in Sanskrit to Hindi have been built for training the system. The system is evaluated using manual evaluation method and geometric average score. Bleu (Bilingual Evaluation Understudy) score refers to the method used for automatic evaluation of Machine translation system. Microsoft Translator hub is used to calculate Bleu Score in this research work.

Table of content

List of Figures

List of Tables

CHAPTER 1
INTRODUCTION

1.1 Natural Language Processing

Natural language processing (NLP) is field of artificial intelligence which uses computerized approach for understanding and generation of human languages, written or oral. NLP found its origin in area of computer science, cognitive psychology and linguistic. NLP emphasis on language processing and generation where first stage is analysis of language for producing meaning and generation stage includes representation. Natural language processing gives wide range of implementation in various fields including machine translation, information retrieve, dialogue system, speech recognition, extraction question-answering etc. This research work investigate the area of machine translation (MT), first computer based application related to natural language. The area of MT examine the use of computer software to translate speech or text from one language to another.

1.2 Machine Translation

Machine Translation is area under computational linguistic that investigate the utilization of software and computer programs to translate speech or text/content from one language to other language. Modern society like today consist of multiplicity in terms of languages. Processes like technology development and globalization all around the world have significantly expanded the need of information translation from one language to another. This need can be found in various fields including education, political organizations, entertainment and industry. Evolution of an evident bilingual Machine translation composition for any two regular languages with confined, restricted resources and devices is difficult and demanding work. Machine translation has picked up enthusiasm of numerous investigates because of its applicability in conquering the dialect boundary. Artificial intelligence goes for giving mechanized frameworks in a few areas to lessen the workload of human specialists. MT systems created for any two languages is referred to as bilingual system, or for many languages is called a multilingual translation system. A bilingual structure from particular Source Language (SL) into particular Target Language (TL) is unidirectional and the translation system or can be bidirectional if translation is allowed in either way between the given

languages. Multilingual frameworks on other hand are proposed to be bidirectional. Machine translation frameworks build varies by number of languages targeted or by the utilized approach. There exist many methodologies in machine translation, these differ in terms of analysis of SL, preserving meaning of the sentence and generation of the target text. Many machine translations are made by using rule based and statistical based approach. Machine translation can be comprehensively partitioned into three classifications Rule based Machine translation (RBMT), corpus based Machine translation and hybrid machine translation approaches.

1.2.1 Rule Based Machine Translation

Rule based machine translation is based on grammar rules. It is processed with help of translation rules and dictionaries covering the morphological, syntactic and semantic regularities of each language respectively. Rule based systems are further categorised in three different type's direct, transfer and Interlingua based Machine Translation System.

1.2.2 Corpus Based Machine Translation

Corpus-based machine translation depends on the analysis of bilingual text corpora. They require less human labour and are fully automatic. They are further of two types one being statistical machine translation and other is example based machine translation. Once the software is implemented it can be used for any other language pair with similar structure format.

1.2.3 Hybrid Based Machine Translation

Hybrid based machine translation is developed by combining the merits of both rule based and statistical based machine translation approaches. They yield with better results in terms of quality and performance.

1.3 Machine Translation System Taxonomy

The diverse approximations to the MT issue can be classified utilizing distinctive criteria:

- Depends on input type: speech or text.
- Depending on the application which utilises the translations: Applications are subdivided in four categories- application for translating input into database query, application for producing translation of given input for post-edition stage, application for generating output along with user, full automatic translation systems.

- Depends on the technology used for translation: technologies are rule based, corpus based or hybrid technology. Rule based is further divided in direct, transfer or Interlingua based machine translation. Corpus based can be statistical based or example based machine translation and hybrid is the combination of the already said technologies.

1.4 Challenges in Machine Translation

There are many challenges to be faced while making automatic translation systems due to structure and syntactic differences among the languages. Some of these issues are listed below:

i. Word Order

Word order is refers to the way in which the words are arranged in the sentence. The word order is given the way in which subject(S), object (o) and verb (v) is arranged in sentence. For example English language has subject-verb-object (SVO) arrangement whereas Hindi has subject-object-verb (SOV) arrangement structure. Language like Sanskrit has free word order i.e. can be in any way SOV, SVO etc [17].

ii. Word Sense

One word can be expressed in different ways when translated to another language. Selection of right word while translating is based on context of the sentence [17].

iii. Pronoun Resolution

The problem of unresolved pronominal references may lead to incorrect translation [17].

Example: A dog saw a cow on the road. It started barking on seeing it

"It" in above sentence can refers to either 'dog', 'cow'. While translating this sentence we need to know gender of because gender of verb is dependent on it. If it needs to refer to dog 'it' should be masculine else if needs to refer to cow 'it' should be feminine.

iv. Idiom**s**

When an idiom in sentence fails to convey the right meaning of translated text it causes error and translation obtained is not appropriate therefore idiomatic expression needs to be taken care of while translation [17].

Example: idiom "a piece of cake" refers to very easy. Sentence:

The English test was a piece of cake.

After translation in Hindi: अंग्रेजी परीक्षण का एक टुकडा था केक।

The translated sentence does not preserve the meaning of the original sentence.

v. Lexical Ambiguity

This type of Ambiguity occurs when words or phrase have more than meaning. Example has been demonstrated below where two sentences are using same word but represent different meaning.

Example: I like watching a live match.

Where do you live?

In above examples word "live" refers watching the match where it is being played in first sentence whereas "live" in second sentence refers to reside or dwell.

vi. Structural Ambiguity

Sentences or phrases with more than one underlying structure results in structural ambiguity.

Example: Visiting relatives are boring.

This sentence can be inferred as: It is boring to visit relatives or Relatives who are visiting are boring.

1.5 Need of Machine Translation

Machine Translation systems are of great importance due to globalization and technology development. Below are the points highlighting the need of Machine Translation over the world today.

i. Socio-Political importance of Machine Translation

There are many regions with more than one spoken languages. The superiority of one language over may even result in loss of that language. The vanishing of a particular language will result in diminishing the culture, therefore Machine translation turns into a social and political need. It is likewise a need of associations like the United Nations for whom multilingualism is both an essential rule and a reality of regular day to day existence [34]. MT turn into a need in the scenario of worldwide economy. To achieve progress and success it is important to provide user interfaces in multiple languages for multinational organizations. Human translation is difficult and turns out to be expensive owing to the hiring of talented and skilled persons [34].

ii. Commercial Importance

Machine Translation is a major need in the situation of worldwide economy. Interpretation of user interfaces in various dialects offers progress in multi-national

organizations. Human translation is a very difficult task and turns expensive owing to the hiring of talented and skilled persons [34].

iii. Literary works Coverage by Machine Translation

Availability of literature work and other content in local dialects helps in breaking the language barrier among people. It is a great degree bewildering to consider a MT structure that can decipher literature works from any vernacular into our native language. [35].

iv. MT for Bridging the Digital Divide

MT can conquer technical boundaries. Internet impacts the working of people nowadays. One can discover huge amount of data on Internet. The usage of this data over any significant portion is very vast as maximum amount of data available is in English language. MT helps in crossing over the computerized isolate, by translating e-mails and web pages into the local dialect of a client [35].

v. Machine Translation used for Assistance in Human Translator

Human translators access electronic dictionaries and online translation systems for the process of translation. The translation database allow translator systems to store texts along with their deciphered forms from one language to another. The translator can look for sentences or phrases in one language to extract corresponding translated sentences in the other language required [35].

vi. Machine Translation used for Cross Language Information Retrieval

Need of data recovery frameworks equipped for seeking content in numerous languages in this multilingual environment Cross Language Information Retrieval (CLIR) manages recovering data written in a language different from client's query language. Example, a client represents his question in Punjabi and can extract information/ significant reports written in English language. CLIR makes utilization of Machine Translation in two different ways: firstly it utilizes MT framework to make an interpretation of foreign language records into clients' question language and then makes an interpretation of the client's inquiry into target dialect. Hence, classical information retrieval techniques uses the target language query to obtain target output in desired language [35].

In this way, MT can possibly defeat dialect hindrances and to make communication between users easier and effective.

1.6 Thesis organisation

This ME Thesis is sorted out into six chapters.

Chapter 2 examines the literature survey on Machine Translation approaches and also represent the overview of existing systems which use different translation approaches on Indian languages.

Chapter 3 discuss about the problem statement and gap analysis.

Chapter 4 throws the light on system architecture and implementation of proposed Sanskrit to Hindi statistical Machine Translation step by step.

Chapter 5 discuss the outcome and evaluate the results obtained from the proposed system.

Chapter 6 gives the conclusion and future scope in area of machine translation from proposed framework.

Chapter 2
Literature survey

2.1 Machine Translation Approaches

Machine translation Systems are of two types multilingual or bilingual. Bilingual translation involve two languages i.e. one specific source and target language. Translation from a specified source to target language is called unidirectional else bidirectional if can be done in either way. Multilingual comprise of more than two languages and is usually bidirectional. Machine translation can be comprehensively partitioned into three classifications Rule based Machine translation (RBMT), corpus based Machine translation and hybrid machine translation approaches. RBMT utilizes linguistic theory while corpus based use data theory and hybrid joins the merits of both RBMT and corpus based machine translation approaches. Distinctive machine translation methodologies can be organised as appeared in figure 2.1 underneath:

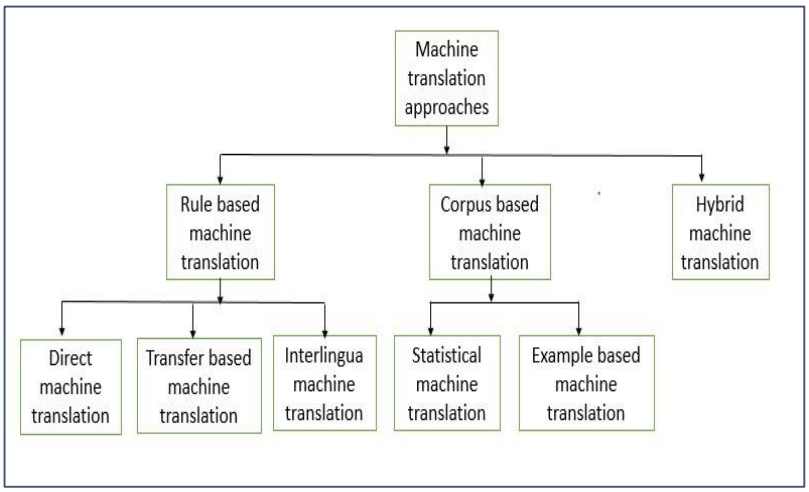


Figure 2.1. Machine Translation Approaches [37]

2.1.1 Rule based approach

Rule based translation comprises multilingual or bilingual lexicon, grammar rules, dictionaries and collection of software programs to practice and apply the rules. These rules are enforced on three stages analysis, transfer, and generation phase. Rules are implied on source language analysis and target language generation in terms of: semantic, syntax, orthographic, morphology and part of speech tagging. The process include association of structure of input text with that of the target text with the help of parser, analyser for source sentence and generator for target sentence along with lexicon transfer which helps in translation. RBMT system always is maintainable and

extensible. The efficiency of the system depends on how efficiency the rules are enforced, there are further three approaches in RBMT i.e. direct, transfer based and Interlingua translation.

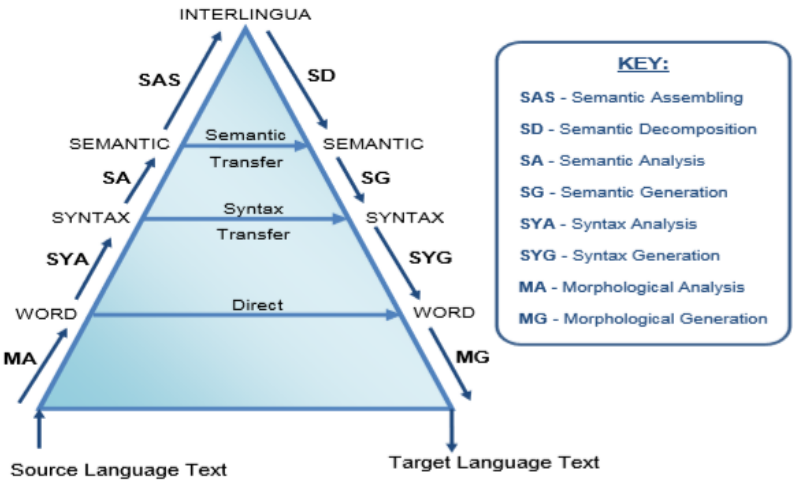


Figure 2.2 Vauquois triangle [33]

The Vauquois triangle shown in figure 2.2 below describes the types of rule based approaches [33]. The vertical height of triangle indicate increase in analysis depth and the horizontal direction indicate the required effort for translation. The base of triangle shows less analysis and more transfer while top shows more analysis and least transfer.

Apertium, an open Source stage to assemble rule based MT frameworks, which gives an MT engine, the encoding of semantic information with help of dictionary and rules files, tools to deal with the provided information and representation used by the engine is compiled along with variety of other tools. They have shown better outcomes of some accessible dialect sets compared with commercial rule based MT systems [1].

2.1.1.1 Direct translation: Direct translation approach follows word to word translation from source to target language which may or may not preserve meaning of the sentence. It is one of the widely used machine translation approach. This approach require high quality and high quantity dictionaries for both the languages (SL and TL) along with some semantic and syntactic arrangement. As shown in figure 2.3 source text is passed as input and starts with morphological analysis which helps in getting the root words from source text. Then is dictionary lookup step for finding corresponding target words for source words. Last step of synantic arrangement helps in rearranging of the words to get best possible output. The yield from this type of approach has many ambiguities

like grammatical errors which can change the meaning of target sentence from source text.

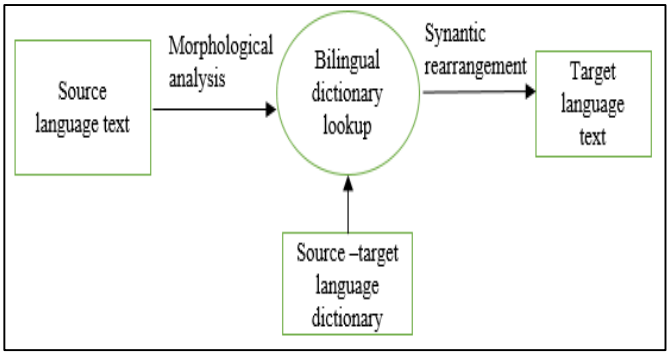


Figure 2.3 Direct Translation [37]

Example: Input (English)- Ajay likes playing football.

Word to word translation - अजय पसंद खेलना फुटबॉल

After synantic arrangement final output - अजय फुटबॉल खेलना पसंद करते हैं

Advantages of Direct MT Approach:

- Effective translation for language pairs with same grammar rules and structure.
- Easy implementation of system

Disadvantages of Direct MT Approach:

- Considers only lexical analysis therefore relationship and structure of words is ignored.
- Can be very costly for multilingual situations and are created for particular language combination
- Sometimes unable to preserve meaning of the source text.

2.1.1.2 Transfer based translation*:* Transfer based approach lies above Direct Approach in Vauquois triangle and consist of three modules analysis, transfer, and generation. In first stage source language parser develop syntactic representation of SL on the basis of syntax, semantics, morphology etc., transfer or second stage uses both bilingual dictionaries and grammatical rules, they transfer the syntactic representation of SL structure to TL structure, and generation stage use morphological analyser to generate target output using target language structure. To develop multilingual system using this approach some components used are analyser component and generator component for each language and transfer component for each pair of language.

Example for a system providing capability of translating five languages it will have five analyzers, five generators and twenty transfer components. Figure 2.4 below describes the workflow of transfer based approach:

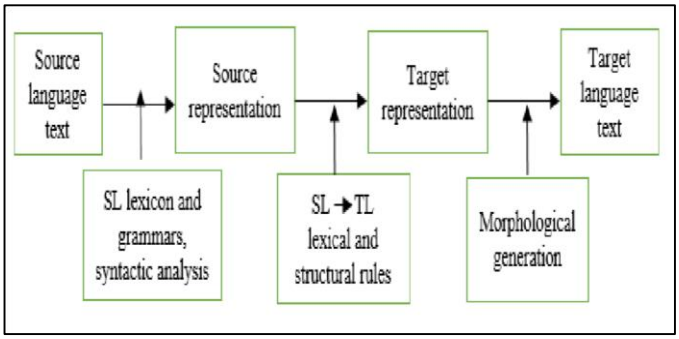


Figure 2.4: Transfer Based Machine Translation [33]

Advantages of Transfer based approach:

- This approach possess modular structure.
- The framework handles ambiguities effortlessly that persist starting with one language to other and has particular structure

Disadvantages of Transfer based Approach:

- Sometimes significance of some source content can be lost in the translation.
- Difficulties in defining and applying rules at all stages while development of the system.
- Difficultly in building transfer module with minimum rules.
- Difficulty in building rules for huge data with numerous dialects and ambiguities.

2.1.3 Interlingua based translation*:* Intermediate translation approach is on top of the Vauquois triangle and has an intermediate language called Interlingua while interpreting translation from source to target dialect/language. . It consist of two stages first being analysis where the system needs to analyse the source language and produce interlingua and second being synthesis phase which uses target language grammars to bringout output from interlingual representation If we have similar language structure then Interlingua language can be utilized for various languages with help of this approach. Interlingua helps in analysis of source data so as to convert its semantic, syntactic and morphological characteristic to provide accurate translation.

This approach is more efficient as it uses linguistic rules for transfer from source to target language. The system uses 2n component for any given n language pair. Figure 2.5 describes workflow for this approach:

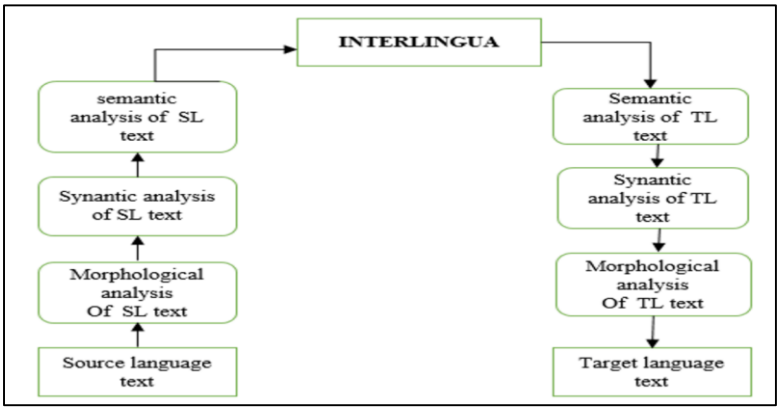


Figure2.5 Interlingua based Machine Translation [37]

Advantages of Interlingua based Approach:

- Gives a meaningful representation and can be utilized in data recovery
- In order to translate to any language from Interlingua system needs to resolve all the ambiguities

Disadvantages of Interlingua based Approach:

- Problem in determining an Interlingua which can preserve the significance of a text.
- Building of Interlingua from source language is very challenging
- Interlingua system has lower time efficiency as compared to Direct MT.

2.2 Corpus based Machine Translation

In Corpus based machine translation source and target language corpus are statistically analysed. This approach uses parallel aligned corpus and is one of the most prominent approach used now a days. Corpus based systems are automatic and require less labour. While making use of this approach in development of any system we need to take care of the available bilingual corpus for the language pair else it will not be possible. It is further categorised in two different methods i.e. statistical and example based machine translation.

2.1.2.1 Statistical machine translation

It is a data driven approach which utilises parallel aligned corpus. It uses mathematical equation for finding probability of source to target language translation [15]. Statistical Machine Translation assigns probability Pr(T|S), where T is target language and S being source input. It uses Bayes' theorem to find maximum probability Pr(T|S), which is written as:

$$\Pr(S|T) = \frac{\Pr(S)\Pr(T|S)}{\Pr(T)} \qquad (2.1)$$

SMT has three stages: language model P(T) for probability calculation of target language, translation model P(T|S) for conditional probability calculation of target to source language and decoder model provides foremost possible translation T by maximizing probabilities that are mentioned earlier and uses search algorithm :

$$T = \text{argmax}\,(P(T|S)) + P(T) \qquad (2.2)$$

Figure 2.6 underneath shows basic parts of SMT system:

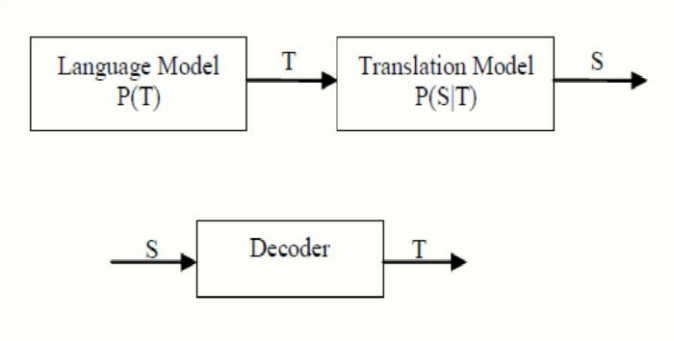


Figure 2.6 Basic Sketch of SMT [17]

It is further classified in three methodologies namely: Word based SMT, Phrase based SMT and hierarchical based SMT.

Word based SMT: Source content is separated to words and afterward translated to target language. Word to word translation is done and objective sentence is obtained by arranging the target words with help of reordering algorithm.

Phrase based SMT: recommended by Koehn separates the source and target languages in phrases. This models is achieved from a word-adjusted parallel corpus according to koehn principle by retrieving phrase pairs those are compatible with word alignment [16]. Antony has proposed some ways of phrase alignment. They have better performance but it degrade for long sentences.

Hierarchical phrases based SMT: Uses combination of both syntax based and phrase based SMT [18].

Language Model P(T)

Fluency of the target text is taken care of in this phase. Language model uses n-gram model to provide the probability of a sentence. It gives probability of single word to the given words that precede it in the sentence and provides likelihood estimation of the sentence. Chain rule helps in decomposing the sentence into product of conditional probability. Given probability of sentence P(S) the probability of sentence with individual words is P(w).

$$\left.\begin{aligned} \mathrm{P(S)} = \ & \mathrm{P(w1, w2, w3, \ldots\ldots, wn)} \\ = \mathrm{P(w1)\ P(w2|w1)\ P(w3, |w1w2)\ P\ (w4|w1w2w3)\ \ldots P(wn|w1w2\ \ldots wn-1))} \end{aligned}\right\} \quad (2.3)$$

Translation Model P(S|T)

The Translation Model compute conditional probability *P(T|S).*It uses bilingual parallel corpus of desired language-pair. This is done by calculating the probabilities of words or phrases broken down from sentences. Therefore for any given sentence there are n number of alignments available. *A(S, T)* denotes alignment set. Let length of target be *l* and length of source be *m* then total of *lm* alignments are possible, therefore P(T|S) is shown as conditional probability $P(T|\,{}^{a}/_{S})$ as shown in equation below:

$$\mathrm{P}(S,T) = sumP\left(S\middle|\,{}^{a}/_{T}\right) \quad (2.4)$$

Decoder

This is last and most important phase of SMT which helps to choose the words with maximum likelihood in translated translation by maximizing probability of translated text.

$$\mathrm{Pr}(S,T) = argmaxP(T)P(S|T) \quad (2.5)$$

Advantages of Statistical machine translation:

• language grammar is not considered and extracts information from corpus also reducing human errors

Disadvantages of Statistical machine translation:

• The translation quality will be exceptionally coarse because of absence of adequate corpora

• It does not function admirably for languages with different word order.

2.1.2.2 Example based Translation

Example based translation, utilizes bilingual corpus, it contains translation memory which stores previously translated text in database if same sentence occurs previous translation is documented. Example based approach saves both user time and processor time for similar data [23]. Example based Machine translation utilizes previous translation for given input. It has three modules i.e. matching module for checking similar data for translation from the dictionary, second is transfer module which helps in translations of source representation to target representation then is the third module recombination where the text chunks retrieved are recombined for whole sentence formation. Figure 2.7 shows the workflow for this approach.

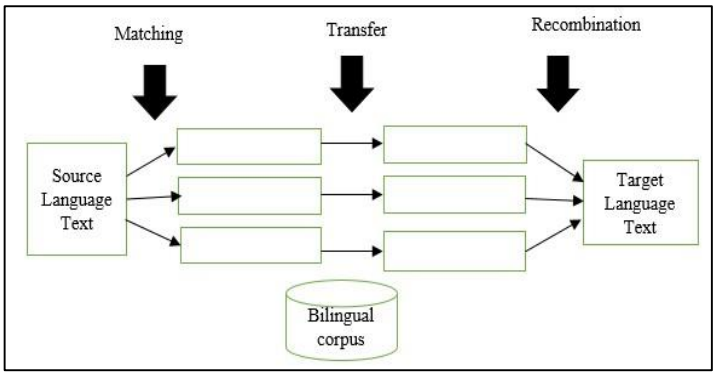


Figure 2.7:Example Based Translation [37]

Advantages of Example Based Machine Translation:

- Established from translation patterns in corpus and retrieve knowledge from corpus

Disadvantages of Example Based Machine Translation:

- EBMT system database needs to have all exclusive examples with all grammatical features so as to provide with accurate translation.
- It is time consuming process of identification of feasible examples.
- Large database refers to more computational for attaining efficiency which is tuff job.
- Sensitive to similarity measure and costly search cost

2.1.3 Hybrid machine translation

Hybrid MT system is derived from merits of both the statistical based and rule based strategies. Rule based require many resources but has high accuracy and statistical require corpora which is sometimes difficult for some language but has high trade off

coverage [27]. Some Indian languages system have been discussed using this approach. Researchers have found this approach to be more accurate [35].

Advantages of hybrid machine translation:

- Exploit features of various approaches and give best possible translation **Disadvantages of hybrid machine translation:**
- Accuracy depends on quality and quantity of bilingual dataset (corpus). Building of high quality corpus is very costly and tedious process.
- This approach fails for language pairs with minimum resources.

2.2 Existing Machine Translation Systems for Indian Languages

ANUSAARAKA

Anusaaraka followed the principle of Direct Translation which intend in translating from one Indian language (source language being Punjabi, English, Kannada, Bengali, Marathi, and Telugu) to other Indian language (target language being Hindi).It used paninian grammar for matching of words [2, 3]. Mapping of local words between source and target language is done. The use of combined architecture of shallow parser and deep parser as described in Apertium and Anusaaraka was used. This project was initiated at IIT Kanpur which later shifted to university of Hyderabad, centre of Applied Linguistics and Translation Studies.

Hindi to Punjabi machine translation system

This system followed Direct Translation Approach since this approach can be used for similar language structure pairs like Hindi and Punjabi. The system was developed in Punjab University, Patiala. The system helps in translation of web pages, text file from Punjabi to Hindi and gives accuracy report of 92%. [5]. The system was later deployed on web.

Mantra

English to Hindi MT system followed transfer based approach and was developed by university of Pennsylvania based on lexicalized tree adjoining grammar (LTAG) used for grammar representation and tree transfer used for translation. Many other tools named entity recognizer, phrase marker, spell and grammatical checker were used. Parsing is done using Earley's bottom up algorithm. The system works well for many other language pairs as well. The system is used in office orders, circulars and azette

notifications and was also tested on administrative documents like notification, appointment letters issued in central government. [6]

SHAKTI

Shakti used transfer based approach and was developed in 2003 for translation from English to Marathi, Telugu and Hindi [7]. It uses nine modules for analysing source language, 24 for executing bilingual tasks, and rest for generation of target language. System uses RBMT and statistical based approach [14].

MAT

MAT was created in 2002, at university of Hyderabad, RC-ILTS, an English-Kannada Machine aided Translation framework which used transfer based approach. The translation system apply UCSG (universal clause structure grammar formalism, it involve post-editing and works at sentence level [8].

MaTra

MaTra followed transfer-based approach and uses frame like structured representation to sort disambiguation words and give accurate translation. It used heuristics and rule-bases to resolve ambiguities. Input analysis is done with help of human assistance and uses sentence splitter for breaking down complex English sentence, then these sentences so obtained are analysed and translated to Hindi. The system is used by editors, translators. Annual reports, news and technical phrases are used as domain of the system [9].

SAMPARK

Another system which follows transfer based approach is SAMPARK which utilizes computational paninian grammar for language analysis and then collaborating it with machine learning. SAMPARK helps in translation of 9 bidirectional pairs for Indian languages. It uses dictionary based and rule based algorithms. [10].

ANGLABHARTI

ANGLABHARTI, a machine assisted system for English to various Indian languages translation was built using Pseudo-interlingua approach which made translation feasible through text generation. AnglaHindi for English-Hindi translation using modules of AnglaBharti was developed in 2002. The verb and noun phrases are translated using abstracted example-base, and hence attained accuracy of 90% [12][29].

UNL-based English-Hindi MT System

MT system for language pair English-Hindi and Marathi based on UNL (Universal Networking Language) was created by IIT, Mumbai which follows Interlingua approach. Source data is changed over in to hyper graph [13].Translation system used pictorial knowledge representation which serves as Interlingua for conversion to target language. The source language is converted into UNL using an enconverter‘, and then converted into the target language using a deconverter‘[29].

Sinhala-Tamil language MT

Sinhala-Tamil language MT was developed in 2011, it used Statistical Machine Translation approach which proposed Morphological Processing and Rule-based word reordering for English-Malayalam SMT and demonstrated increase in accuracy with the help of morphological information and word ordering of source and target languages [19].

English to Hindi MT

An English to Hindi MT framework by joining RBMT and phrase based SMT approach was created at IIIT Hyderabad in 2010 .Source analyser use brill’s POS tagger for linguistic analysis and then is converted to chunk based dependency tree. Next stage performs reordering by transfer grammar for translation [20].

Google Translate

Google Translate used detect patterns for finding best possible translation in millions of documents and used SMT approach. It translates documents, text, web pages. Later Google developed neural machine translation using recurrent neural networks with help of attention mechanism and LSTM encoder-decoder model which helped in increasing accuracy, speed, robustness [21].

EnTel

“EnTel”, a SMT based English to Telugu system described by Johns Hopkins University Open Source Architecture (JOSHUA) which has used dictionary proposed by Philip brown for language pair English-Telugu for training [22].

ANUBAAD

ANUBAAD (2004), a translation system used for headlines translation from English to Bengali. First exact match is found from memory, if not found generalised tagged

example-base is used to translate still if system is unable to translate phrasal example base is used. Finally if all fails then heuristic translation strategy is executed [24].

VAASAAANUBAADA

VAASAAANUBAADA, Bengali-Assamese translation system was developed in 2002 for News Texts using EBMT approach. It comprised pre-processing and post processing tasks, and used Pseudo code for aligning bilingual corpus. Long sentences were fragmented and if match no found backtracking techniques are used [25].

SHIVA

SHIVA MT system is another illustration of example based approach developed by Indian Institute of Science, Bangalore, and IIIT, Hyderabad, along with Carnegie Mellon University. The system was used for feedbacks and trails and with the feedback received performance for the system was developed [26].

Anuvadaksh

Anuvadaksh which helps in translation from English to six different Indian languages (Marathi, Oriya, Tamil, Urdu, Hindi and Bangla) machine translation system. It is a consortium based project uses four technologies: MT based on Tree-adjoining Grammar, develop MT after analysing and generating rules (anlagen), statistical based translation (SMT), example based machine translation [14, 28]. Bengali to Hindi machine translation uses contextual information from Hindi corpora for better lexical translation [29, 30]. It used lattice based data structure for translation and is evaluated using BLEU score [31].

SaHiT

Sanskrit-Hindi Machine Translation system (SaHiT) which follows statistical approach. The system was trained on two platforms: MT Hub and Moses [36], and discusses the errors on MT hub. The system uses 24K parallel and 25k monolingual sentences for training and obtains BLEU scores of 41 and above [32].

EtranS

English-Sanskrit machine translation system adopts statistical approach which possess context free grammar technique. It has two components mainly parser and generator component where parser analyse the sentence and give grammatical information and generator carries this grammatical information to produce translation [33].

The table 2.1 describes various existing machine translation systems for Indian language pair. It represents system name, year of development, approach used in building the system, language pair used and features of the developed system.

Table 2.1: Existing Machine Translation System on Indian Languages

System	Year	Approach	Language pair	Features
Anuvadaksh	2016	Hybrid	English to Indian languages	Use four technologies :SMT, EBMT, MT based on Tree-adjoining Grammar, analyse and generate MT system based on rules (anlagen)
Google NMT	2016	Statistical	Various language pairs	using recurrent neural networks with help of attention mechanism and LSTM encoder-decoder model
Sinhala Tamil	2011	Statistical	English and Sinhala/Malayalam	Suggested Rule-based word Reordering and Morphological Processing
enTel	2011	Statistical	English to Telugu	Uses JOSHUA
Bengali Hindi machine translation system	2011	Hybrid	Bengali to Hindi	Lattice based integrated with transfer based
Web developed Punjabi to Hindi MT	2010	Direct	Hindi to Punjabi	Deployment of the proposed Hindi to Punjabi translation system on web
English Hindi machine translation	2010	Statistical	English-Hindi	Uses brill's POS tagger for linguistic analysis, then converted to chunk based dependency tree and reordering by transfer grammar
MaTra	2008	Transfer based	English-Hindi	Build on intuitive intermediate representation with simplification of parsing algorithm
Punjabi to Hindi Machine translator	2007 2008	Direct	Punjabi - Hindi	Requires post processing and achieved accuracy of 90.7%
Sampark	2005	Transfer based	9 bidirectional pairs	Merge with machine learning and uses computational paninian grammar for analysing language
ANUBAAD	2004	Example based	English to Bengali	Uses Example-based, Phrasal example-based, Generalized Tagged example-based or heuristic translation strategy is used for headlines translation

SHAKTI	2003	Transfer based	English and Indian Language	Linguistic rule base with Statistical processing
AnglaHindi	2003	Interlingua	English to Hindi	Pseudo interlingual rule-based, verb and noun phrases use abstracted example-base for translation
SHIVA	2003	Example based	Hindi to English	Uses statistical approach and linguistic rules to infer linguistic information.
MAT(English -kannada)	2002	Transfer based	English and Kannada	morphological analyser, Universal Clause Grammar Structure & post editing
VAASAAN UBAADA	2002	Example based	Bengali – Assamese	Pre-processing and post processing task, backtracking when not matched.
ANGLABH ARTI	2001	Interlingua	English to Indian language	Uses PLIL (Pseudo Lingua for Indian Languages)
English-Hindi MT System based on UNL	2001	Interlingua	English-Hindi, UNL - Hindi, Hindi – UNL	Universal Networking Language as Interlingua
Mantra	1999	Transfer based	English and Hindi	Uses a Synchronous TAG, TAG and Tree Transfer Output based on analysis of the input text with help of a bilingual dictionary and full parser.
ANUSAAR AKA	1995	Direct	Kannada, Bengali, Marathi Telugu, Punjabi to Hindi	Uses Paninian grammar and matches words between SL and TL.

2.3 Conclusion

We have studied different Machine Translation methodologies and existing Translation systems on Indian languages which have utilized distinctive formalisms according to their application. It is concluded that transfer based systems can be stretched out to dialect matches in a multilingual situation and are more flexible. Most Indian systems follow hybrid and statistical based approach. The reason being that rule based systems fails to give high accuracy because Indian languages contain many dialects and are morphologically rich. Less quantity of corpora is available in many multilingual language pairs and also very less systems are available for south Indian languages. The machine translation systems so far created have numerous inadequacies as rule set, word reference (dictionary or corpus availability), and translation approach. Another future motive is attainment of multilingual framework system with reduction in corpus preparation burden and also demands some innovative translation approach to be developed. It is noticeable from the survey undertaken that further work is needed in the field of MT in order to deliver intelligible translations.

Chapter 3
Problem Domain

World is turning in global town by each passing day. Every country has its own official language and these languages are according to their geographical and cultural differences so we need machine translation system for communication across the globe

3.1 Problem Gap

We have studied different Machine Translation methodologies and existing Translation systems. The machine translation systems created so far have numerous inadequacies as rule set, word reference (dictionary or corpus availability), and translation approach. Languages like Sanskrit which possess free word order and grammatically rich are difficult to translate. English language is used for enormous amount of text available in digital format. We have a lot of literature works done in Sanskrit, therefore there is need for translation system which helps us with this language. We have systems for English to Sanskrit language conversion. There is large amount of population who cannot comprehend Sanskrit and English language therefore we need a Translation system for Sanskrit-Hindi language pair (Hindi is official language of India and known by many). So we are where we have Sanskrit as source language and Hindi as target language using statistical building a system Machine Translation approach.

3.2 Objective

The Sanskrit to Hindi Statistical Machine Translation system has following objectives:

- Building of bilingual corpus of Sanskrit-Hindi language pair.
- Studying the grammar structure of both the languages.
- To learn about Thot tool for building of statistical Machine Translation.
- To study the different models i.e. language model, translation model and search algorithm used in building up the proposed System.
- To prepare data and learning the working of training pipeline used in Thot.
- To test and evaluate the results

3.3 Methodology

Sanskrit to Hindi statistical Machine Translation system is build using Thot. The system uses n-gram model for language model, phrase based translation for translation model

and branch and bond search algorithm. The system is deployed on cloud using Virtual Private Server. A parallel corpus of 1000 sentences in Sanskrit to Hindi have been built for training the system. The system is evaluated using manual evaluation method and geometric average score. Bleu is calculated with help of MT hub.

3.4 Conclusion

Research gap analysis obtained by studying various existing translation systems helps in understanding different approaches used for development of translation model. Review of previous work done in the area of Machine Translation helped us in problem formulation and choosing the desired language pair (Sanskrit to Hindi) as well. The language pair is so chosen since we have large amount of literature work done in this language but very less amount of work done for the said pair. This identification led to the formulation of problem statement. In this chapter, the major objectives were discussed. In the next chapter, the proposed model is discussed in detail.

Chapter 4

Implementation of Statistical Machine Translation

This chapter presents, design and implementation of System, working steps included in order to build Sanskrit to Hindi statistical machine translation system. This includes development of corpus, installation of Thot, Architecture of system, training pipeline in the system and deployment of system using Virtual Private Server.

4.1 Development of Corpus

Statistical based Machine Translation approach is corpus based i.e. it needs bilingual parallel corpus of source-target language pairs. The system proposed uses Sanskrit as source and Hindi as target language. A parallel corpus of 1000 sentences have been developed manually since dataset for this language pair is not available online. The dataset/corpus available are for English to Sanskrit or English to Hindi but no direct dataset for Sanskrit to Hindi language exist. Some paragraphs and sentences have been considered for development of corpus. The data is obtained from some sentences given in avainble Sanskrit with their meaning in Hindi and by using English as intermediate language.

4.1.1. Structure of Sanskrit and Hindi language

Sanskrit is an ancient language whose formation has not been changed since its commencement and is called mother of many languages. Puran, Upanishads and Veda are scripted in Sanskrit language. Sanskrit uses concept of root word called “dhatu” which helps in deriving new words from basic words. Sanskrit language is elaborative complete in its sense.

Table 4.1Comparison between Sanskrit and Hindi language

	Sanskrit	Hindi
Era	Classical language (very common in ancient times)	Commonly used nowadays
Types of noun	Sanskrit has least twenty-one forms depending on number and vibhakti	Hindi has five forms of noun
Root word	Uses root word for deriving new words “dhatu”	No root words some words are derived from Sanskrit language.
Word order	Free word order	Subject object verb

Tenses	Tigant kriya and kridant kriya	Skarmak kriya and askarmak kriya

4.1.2 Corpus Partition

The corpus is portioned in three parts before training and tuning the system. It is divided in three parts mainly training set, development set, test set.

Training set: used for training different statistical models used in translation. Consist of thousands of sentences. In our system we have used 1000 Sanskrit- Hindi parallel corpus sentences to train the system.

Development set: used for tuning of the system, consist few sentences from the corpus. For development of the proposed system 70% of the dataset is used in development set Test set: used to test the translation performed by the system. They are used for evaluating target sentences with reference sentences. It is composed of few sentences.

The proposed system uses 30% of the dataset for testing set.

Table 4.2 Corpus Partition

Bilingual Sanskrit to Hindi dataset	1000 sentences
Training set	1000 sentences (100%)
Development set	700 sentences (70%)
Test set	300 sentences (30%)

4.2 Installation of Thot

Thot is an open source toolkit which uses statistical machine translation approach for building of translation system. The tool helps to train phrase based translation models. It includes phrase based translation decoder, uses n-gram language model, EM incremental algorithm for word alignment, branch and bound algorithm for searching. The code for Thot is available on github. In order to install Thot, we need to install some tools that are automake, autoconf and libtool packages in Ubuntu platform. Thot tool can be installed on windows and Mac systems as well. In order to install it on Windows Cygwin environment needs to be installed and if needed to be installed on Mac, MacPortsc are used. After we have installed the tools needed we can start with installation of Thot. For installing tools in Ubuntu we run following commands:

Sudo apt –get install autotools-dev

Once installation of autotools have been done, proceed with Thot installation as given in the steps below:

i. Get Thot package from:

$ git clone https://github.com/daormar/thot.git

ii. "cd" command is used to direct to the source code package and type "./reconf".

iii. "/configure" is used to configure the package.

iv. For compiling the package type "make".

v. "make install" helps in installing programs.

vi. For removing object files and program binaries from source code type clean.

In order to give a specific path for installing Thot the command is given below:

$ configure --prefix=<absolute-installation-path>

vii. For ending the installation process, Thot is added to system PATH by executing following commands:

$THOT_HOME_DIR=<absolute-installation-path>$export

PATH=$PATH :${ THOT_HOME_DIR}/bin

4.3 Architecture of the system

The architecture plays central role in building the SMT system. Components of architecture includes pre-processing, Translation Model (TM), Language Model (LM), and decoder of SMT. Figure 4.1 depicts the architecture of the proposed system used for developing Sanskrit to Hindi translation system.

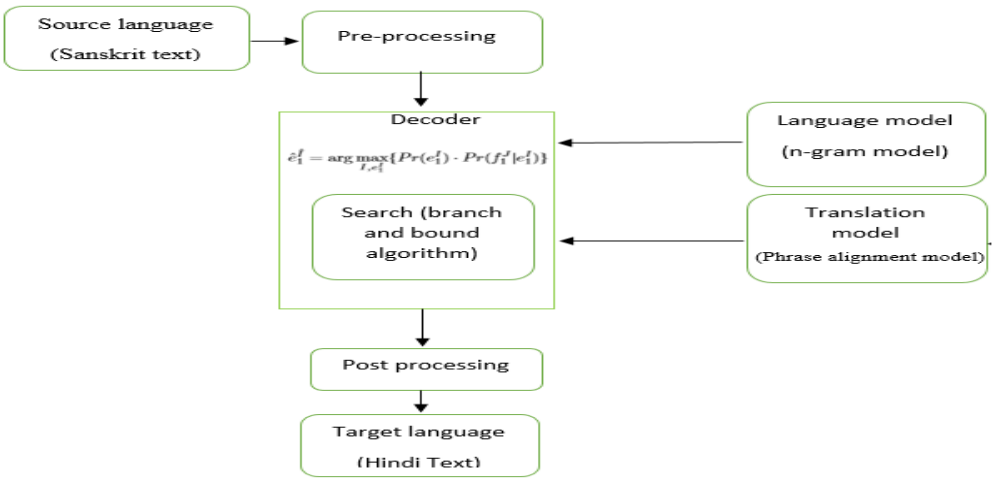


Figure 4.1 Architecture of Proposed System [37]

Statistical Machine Translation (SMT) uses mathematical equation for finding probability of source to target language translation. Given source language F, for the source sentence f_1^J we want equivalent target sentence e_1^I in target language E. From all possible sentence obtained for source sentence we want the target sentence with highest probability according to following equation by applying Bayes decision rule:

$$\hat{e}_1^I = \arg\max\{\Pr(e_1^I).\Pr(f_1^J|e_1^I)\} \quad (4.1)$$

Where $\Pr(e_1^I)$ is modelled by language model and $\Pr(f_1^J|e_1^I)$ is modelled by translation model. The basic problems in the field of SMT is training and search problem. The SMT consist of language model to take care of fluency in model, data in target language is used. Language model uses n-gram model for finding probability of the sentence. Translation model uses both the data in source as well as in target language and deals with lexical correspondence between the languages. Phrase based alignment models are used .Search problem is solved using branch and bound algorithm.

4.3.1 Preliminaries in Statistical machine translation system:

4.3.1.1. Language model (n-gram model)

This model measures the well-formedness of the sentence e_1^I as a sentence of the language E. Language model uses n-gram model to provide the probability of a sentence. It gives probability of single word with respect to the given words preceding to it in the sentence. Goal is to estimation of likelihood of sentence. Chain rule helps in decomposing the sentence into product of conditional probability. Given probability of sentence P(S), P(w) gives probability of sentence with individual words.

$P(S) = P(w_1, w_2, w_3, ..., w_n)$ (4.2)
$=P(w_1)\ P(w_2|w_1)\ P(w_3,|w_1w_2)\ P\ (w_4|w_1w_2w_3)\ldots P(w_n|w_1w_{2\ldots}w_{n-1}))$

According to bigram models the probability of any given word is dependent on Preceding word.

$$\Pr(x) = \prod_{i=1}^{|x|} p(x_i|x_1 \ldots x_{i-1}) \approx \prod_{i=1}^{|x|} p(x_i|x_{i-1}) \quad (4.3)$$

Where $p(x_i|x_{i-1})$ determines frequency of $x_i$word and x_{i-1} is the previous word. Calculation of count of occurrences in training text is normalized to:

$$p(x_i|x_{i-1}) = \frac{\mathrm{c}(x_i - x_1^i)}{\sum x_i \, \mathrm{c}(x_i - x_1^i)} \tag{4.4}$$

Where $\mathrm{c}(x_i)$ refers to number of occurrences of x_i in source text.

When n > 2, probability is calculated on last n-1 words. Here x_{-n+2} gives beginning of sentence and $|x| + 1$ end of sentence. Therefore the equation set is as shown below:

$$p(x) = \prod_{i=1}^{|x|+1} p(x_i|x_{i-n+1} \dots x_{i-2x_{i-1}}) \tag{4.5}$$

The general equation for n gram model is:

$$p(x_i|x_{i-n+1}^{i-1}) = \frac{\mathrm{c}(x_{i-n+1}^{i-1})}{\sum x_i \, \mathrm{c}(x_{i-n+1}^{i-1})} \tag{4.6}$$

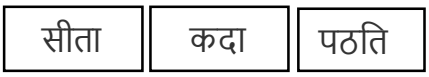
Where x_{i-n+1}^{i-1} is source sentence with $i - n + 1$ as start and $i - 1$as end of sentence.

In Thot tool n=3 is used in case of n gram model and probability is calculated as shown later in Figure 4.9.

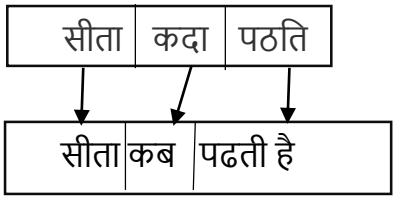
4.3.1.2. Translation model (phrase alignment model)

The translation model measures how good f_1^J is for a given translation of e_1^i . Phrase translation models is general way to for implementing translation models. These models follow three steps under generative point of view:

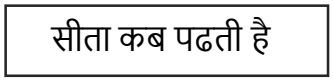
i. Source sentence is divided into segments.



ii. Target translation for each segment.



iii. Reordering of the obtained translated text for given source sentence to get complete translated output.



In Phrase Based Translation, it is accepted that the relationship between source and target word sentences is explained by mode of hidden variable $\tilde{a} = \tilde{a}_1^k$. The Phrase Based alignment model is used for phrase based translation in this tool.

Here the translation process is achieved by bisegmentation of source and target sentences. Consider, bisegmentation of length K of sentence pair $(f_1^j|e_1^i)$, $\tilde{A}(f_1^j|e_1^i)$,is defined in terms of triplets as $(\tilde{f}_1^k, \tilde{e}_1^k, \tilde{a}_1^k)$ where $\tilde{a}_1^k$ is hidden variable used for showing one to one mapping between k phrases of both source and target sentences. Bisegmentation can be seen as phrase-level alignment between two sentence pairs. The hidden variable allows to regenerate the probability distribution $\Pr(f_1^j|e_1^i)$ without any loss as given below:

$$\Pr\left(f_1^J \middle| e_1^I\right) = \sum_{a_1^k} \Pr\left(\tilde{a}_1^k, \tilde{f}_1^k, \tilde{e}_1^k\right) = \sum_{\tilde{a}_1^k} \Pr\left(\tilde{a}_1^k | \tilde{e}_1^k\right) . \Pr(\tilde{f}_1^k | \tilde{a}_1^k , \tilde{e}_1^k) \qquad (4.7)$$

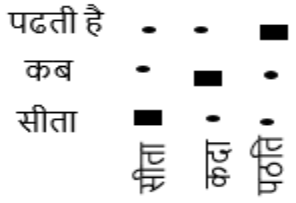
Phrase based translation depends on bilingual phrases which are obtained previously. There are three ways of getting these phrases from corpus:

1. Word based alignments
2. Syntactic phrases
3. Sentence alignments by joint probability

Thot toolkit uses first approach of word based alignment. The extraction of phrases is done with help of additional constraint i.e. consistency of the bilingual phrase with word alignment matrix A as demonstrated in equation below:

$$BP\left(f_1^J, e_1^I, A\right) = \{\left(f_j^{j+m}, e_i^{i+n} : \forall(i', j')A : j \leq j' \leq j+m \leftrightarrow i \leq i' \leq i+n\right\} \qquad (4.8)$$

The below figure shows the example of word alignment matrix with bilingual phrases.



Source phrase	Target phrase
सीता	सीता
कदा	कब
पठति	पढती है

Figure 4.2 Word Alignment Matrix (left) and corresponding Bilingual Phrases (right) [38]

Thot toolkit provide operations for better alignments. The following operations are incorporated in the tool used.

Union: acquire union for two given matrices.

Intersection: acquire intersection for two matrices.

Sum: acquire sum for two or more matrices.

Symmetrisation: acquire "something" in between union and intersection of two matrices.

4.3.1.3. Search (branch and bound algorithm)

The search problem in SMT is used to find target translation e_1^i for any given source sentence f_1^j with highest probability. Branch and bond algorithm is used for searching. Figure 4.3 gives flowchart for the proposed search algorithm. It includes following steps:

1. Initialise the stack with the null hypothesis
2. hypothesis with highest score is removed from the stack
3. If this hypothesis obtained is objective hypothesis, retrieve it and terminate
4. else produce extensions for the given hypothesis and push them into the stack
5. Go back to step 2

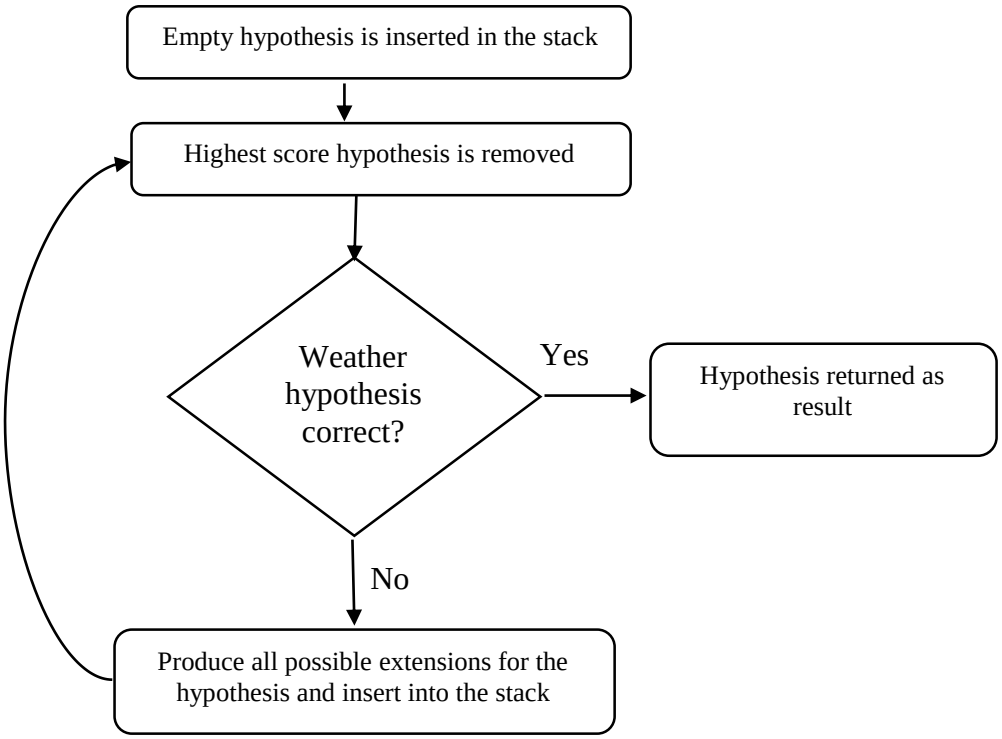


Figure 4.3 Search Flowchart [39]

The pseudo code for algorithm of proposed Branch and bond search algorithm is shown in figure 4.3 below. The algorithm uses stack data structure. The algorithm keeps expanding until a complete hypothesis is found. Hypothesis is said to be complete if all source positions are aligned, $SP_h = 0$. Partial translation for any given hypothesis is given by obtain trg sent function. The last element in given vector used to represent hypothesis is returned using Back function. The hypothesis are inserted in stack by means of push function and top is extracted with help of pop function. If hypothesis are not complete in top of stack expand function is used. Set H is used to store expansion results. Every hypothesis in H is assigned score, calculation of this score is done by incrementally from predecessor hypothesis score h and also from the information obtained from last extension. The expansion algorithm gives the alignment of the source position $(j_{1,}, j_2)$, from set PP(SP_h) where PP() gives set of word positions, the set of possible phrase positions obtained from word positions. Scoring function is calculated $q(h)=f(h) + r(h)$ where $f(h)$ provides logarithm of probability of h and $r(h)$ is o. once search completes these variables helps in retrieving target sentence with highest probability. The following are the abbreviations used in the algorithm:

i. SP represents the set of currently unaligned positions of the source sentence.

ii. m represents the number of target words that compose the partial translation, e_1^m of the source sentence.

iii. σ represents the last n − 1 target words that has been added to the partial translation e_1^m, where n is the order of the n-gram language model. In other words, σ is the language model history of the current partial hypothesis.

iv. j represents the rightmost source position of the last source phrase that has been aligned. j is required to appropriately generate distortion probabilities.

iv. BOS refers to begin of sentence

vi. EOS refers to end of sentence

Pseudo code for the explained above algorithm is given in algorithm shown below:

Input : f_1^J, n (order of n gram language)
: $p(e_1^i)$ (n gram language model)
Output : $\tilde{e}_1^I$ (optimal translation)
Auxiliary: Stack
Begin
$h_\emptyset$:= [((0,0), BOS)]
push $(s, 0, h_\emptyset)$
end :=false
while !end **do**
(q, h) :=pop(s)
if $SP_h = \emptyset$ then
$\tilde{e}_1^I$:=obtain_trg_sent(h)
end :=true
else
$'H$ =expand $(h, T_{j1,j2})$
$e_1^{m\prime}$:= obtain_trg_sent(h)
$m' := |e_1^{m\prime}|$
σ' := $(n-1, BOS\ e_1^{m\prime})$
j'':= back(h)
for all $h' \in H\ do$
$((j', j), \tilde{e})$:=back(h')
$m := |\tilde{e}| + m'$
$p := \prod_{i=m'+1}^{m} p(e_i|e_{i-n+1}^{i-1})$
$p(m|m'+1).p(j'-j'')$
$p(j-j'+1|m, m'+1).p(f_1^j|\tilde{e})$
if $SP_{h'} = \emptyset$ then
$p := p.p\left(EOS\middle|tail(n-1, \sigma'\tilde{e})\right).p(J|m)$
$q' := q + \log p$
push(s, q', h')
end

Algorithm 1 Branch and Bond Search [41]

4.4 SMT pipeline Using Thot

The statistical Machine Translation (SMT) using Thot tool follows a series of steps to train the system for building any translation system. Figure 4.4 depicts Statistical Machine Translation pipeline with commands showing steps to be followed in building the system.

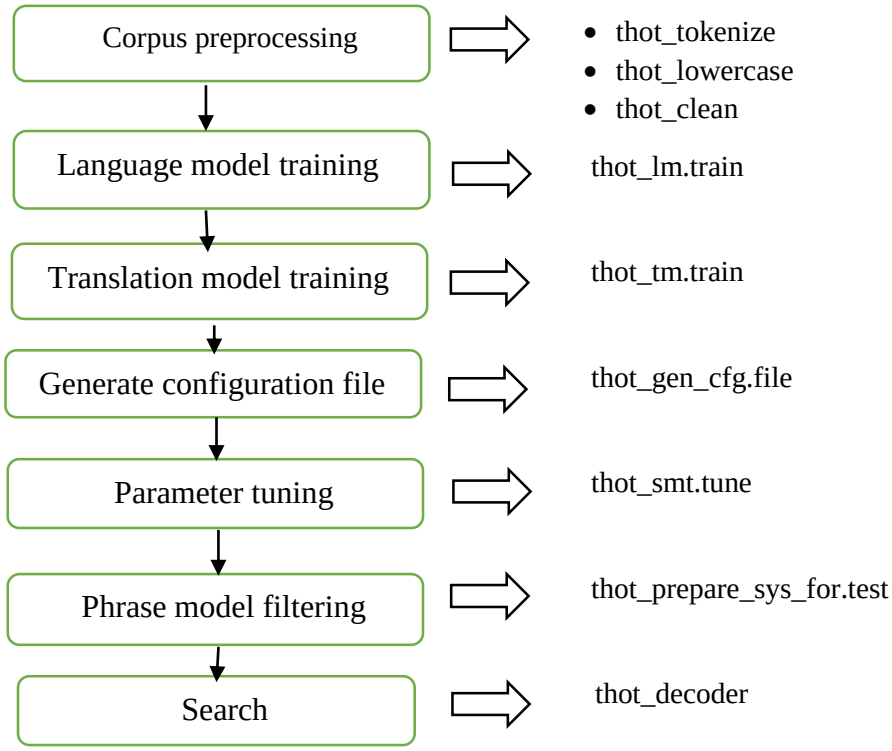


Figure 4.4 SMT pipeline [39]

4.4.1 Corpus Preprocessing

Data Preparation involves three main aspects that are tokenizing, lowercasing and cleaning of corpus. Preprocessing is to be done before training the system. For preparation of data we have commands to be executed. And are discussed in brief later

i. Tokenizing the corpus

It is essential to break sentences in corpus in words, symbols etc. called tokens. Thot has thot_tokenize tool for providing this functionality. The tool uses –f parameter to receive the file to be tokenized. Raw corpus is input to this tool and tokenized corpus as output. The figure 4.5 shows the commands to be followed while tokenizing the source and target data, the figure 4.6 shows tokenized data after executing the commands.

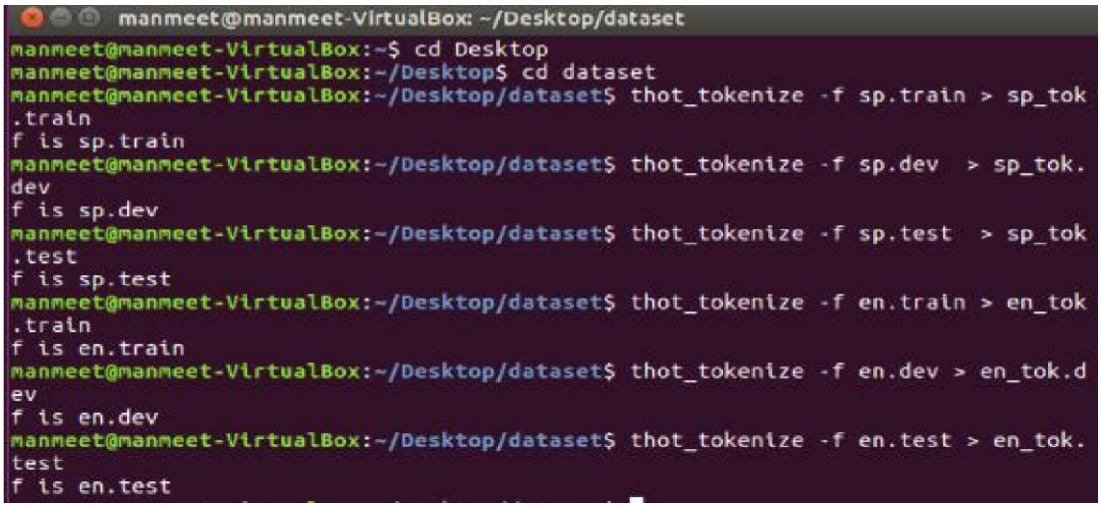


Figure 4.5 Tokenization data

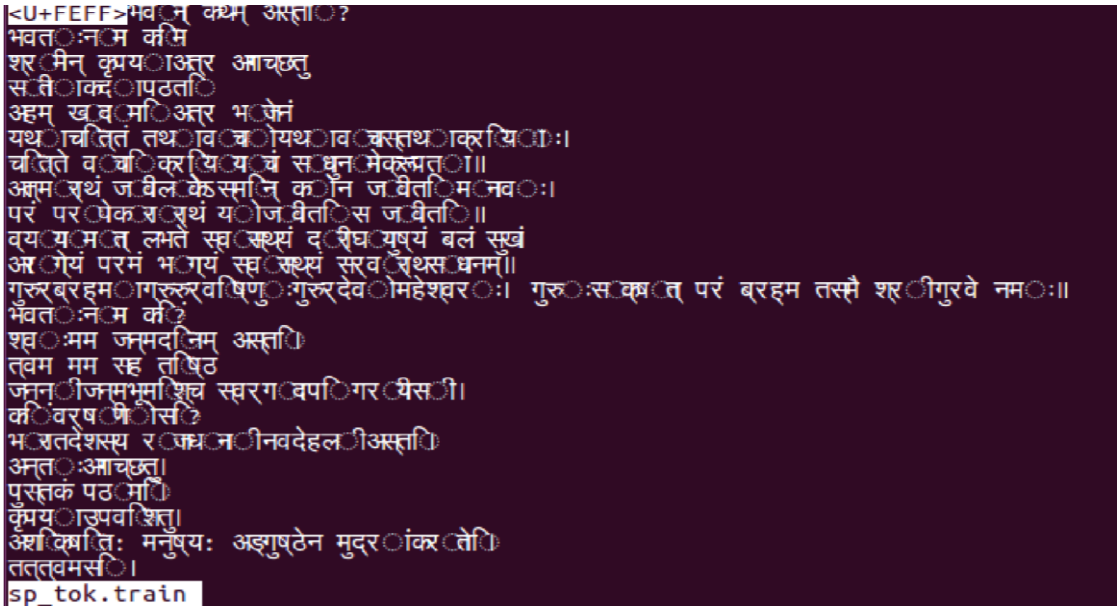


Figure 4.6 Tokenized Data

ii. Lowercasing of Corpus

After corpus tokenization, it is useful to lowercase the data. This helps the system to free itself from appropriate capitalization task of translated task. This will help to better exploit the corpus available for training in order to get better translation quality. Thot tool provide thot_lowercase tool for this purpose. The files to be lowercased are received using –f option. The figure 4.7 shows the commands needed for lowercasing the data and figure 4.8 shows the lowercased data after the commands have been executed.

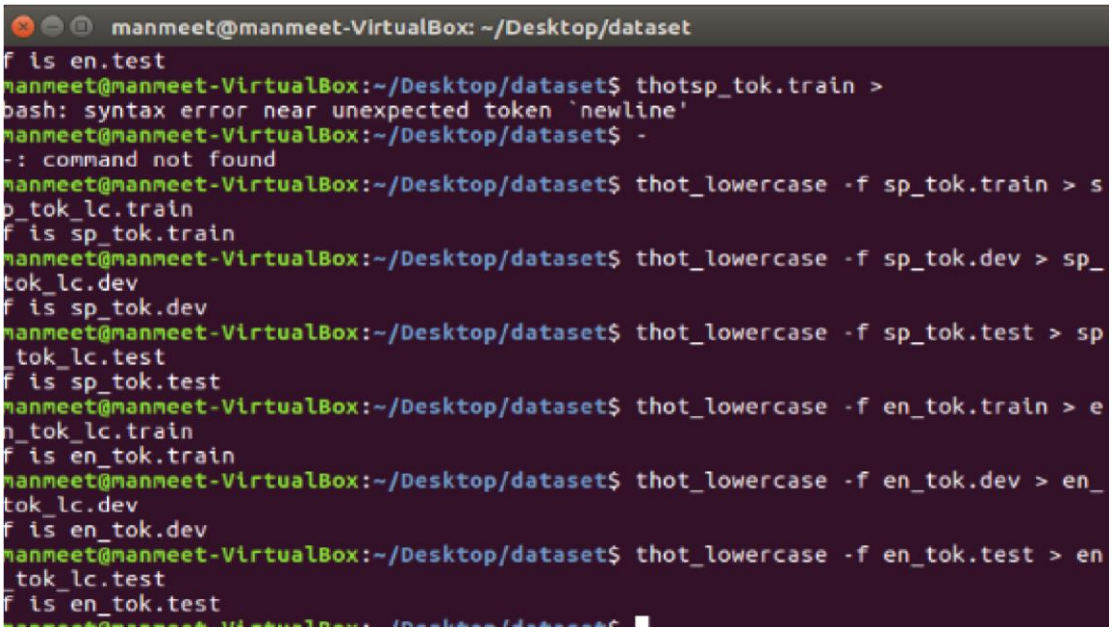


Figure 4.7 lowercasing the data Corpus

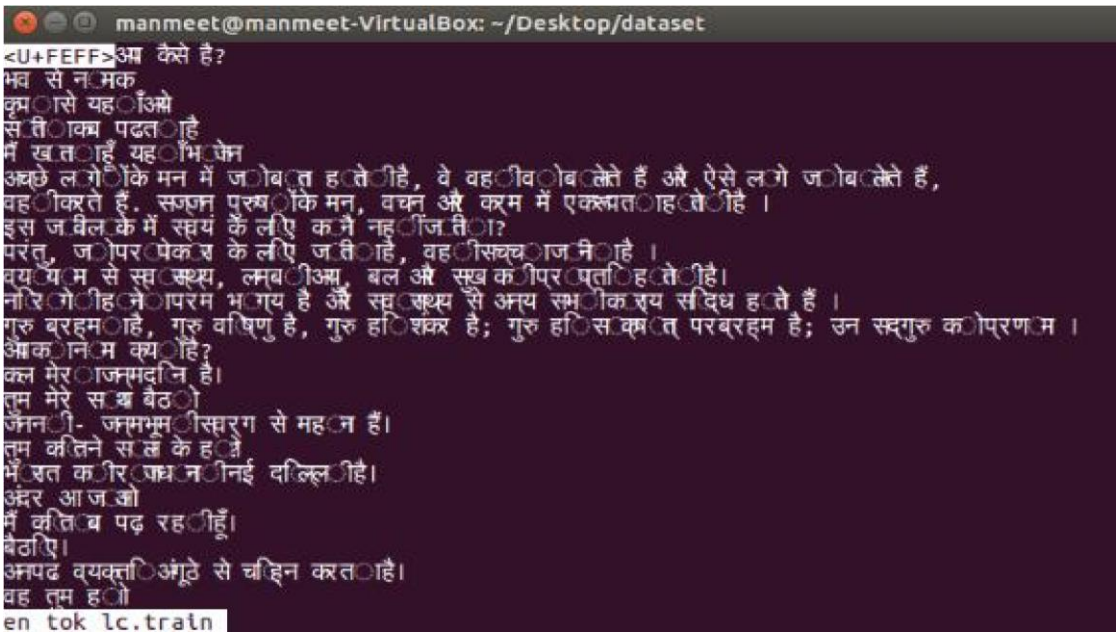


Figure 4.8 Lowercased Data

iii. Cleansing the data

The bilingual parallel corpus needs to be cleaned by omitting extremely large sentence pairs before training else they will increase the parameter estimation time. The sentences with large length may cause segmentation errors. Thot includes thot_clean_corpus_ln tool for this .It contains input parameters as shown in table 4.3. The input parameters consist of source and target sentences file, also determines maximum length allowed in sentence along with output directory.

Table 4.3 Parameters for Cleaning

-s (string)	file with source sentences
-t (string)	file with target sentences
-a (int)	maximum sentence length allowed
-d(int)	Maximum standard deviation allowed in source and target sentences length

Figure 4.9 depicts the commands for cleaning the data and shows that maximum sentence length allowed is 80.It cleans bigger sentences to avoid extra parameter estimation time.

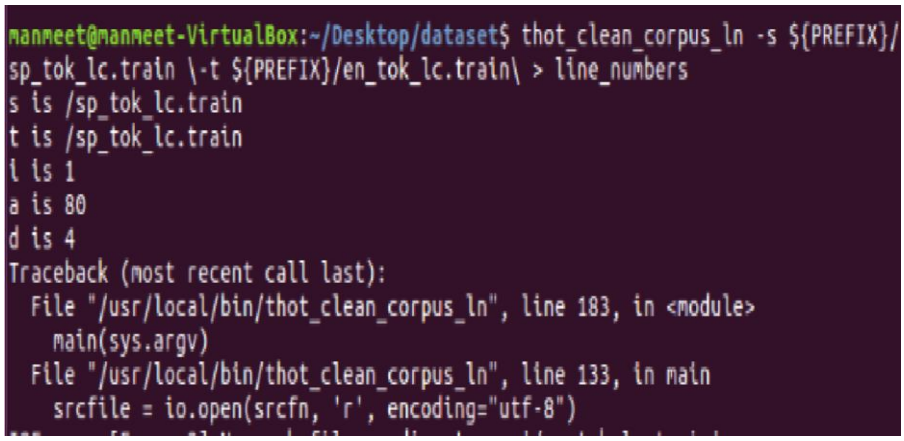


Figure 4.9 Cleaning of Data

4.4.2. Language model training:

Thot provide n-gram model for language model training. It uses thot_lm_train tool to train the language model. The n-gram model used has value n=3. The table 4.4 give list of input parameter used in the command for training the language model which include number of processors, output directory, and order of n gram model needs to be specified in input command.

Table 4.4 Input Parameters for Language Model Training

-pr (int)	Number of processors for estimation performance.
-c(string)	Monolingual corpus used
-o(string)	Output file directory
-n(int)	Order of n-gram model

Figure 4.10 shows the language modelling command to find the probability of word with respect to its preceding words and Figure 4.11 gives n gram values where order of n gram model is 3.

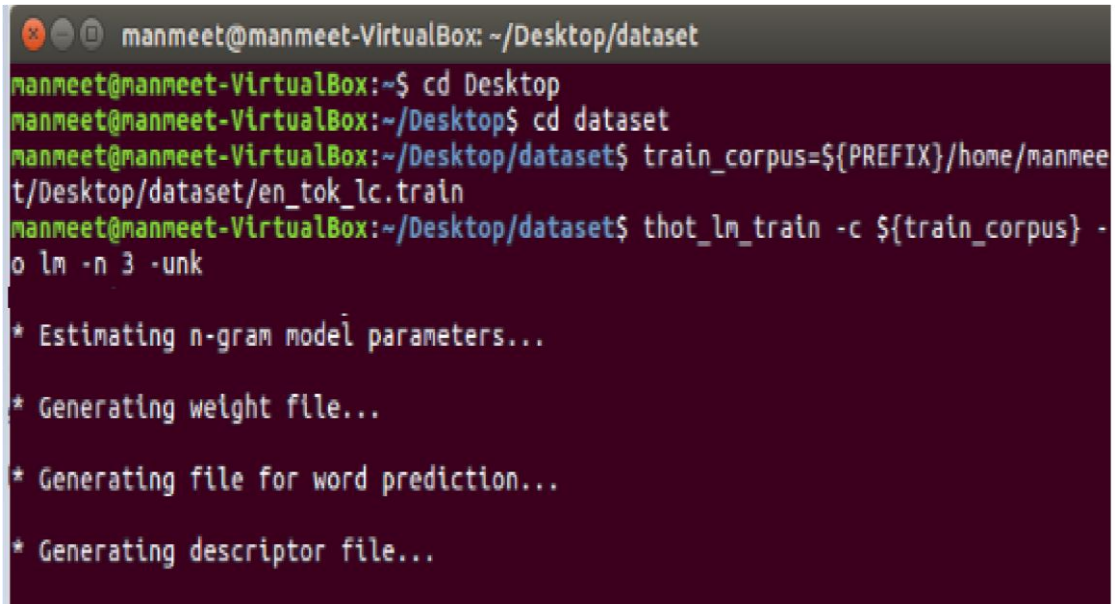


Figure 4.10 Language Model Training

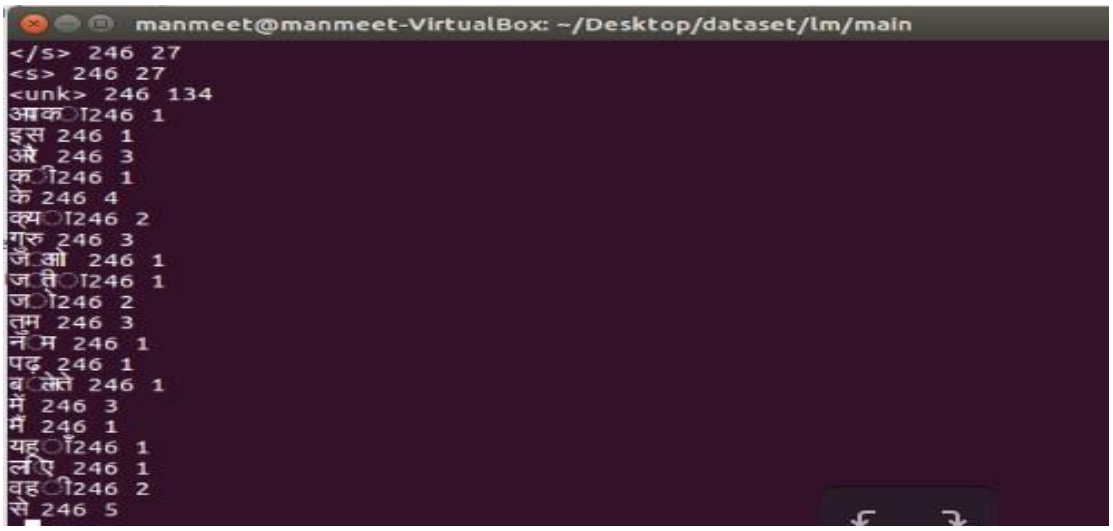


Figure 4.11 n-gram Values

4.4.3 Translation model training

Phrase based models are used for translation model training. Thot uses thot_tm_train tool and below table 4.5 shows the input parameters for this command which include processors used, source and target file along with output directory.

Table 4.5 Parameters for Translation Model Training

-pr (int)	Processors for estimation performance.
-s(string)	Source sentence file
-t(string)	Target sentences file
-o(string)	Output file directory

The translation model provide us with best alignment between source and target sentences and also generates phrase based models. The commands used for training of phrase models and results obtained for the proposed system is shown in figure 4.12 below

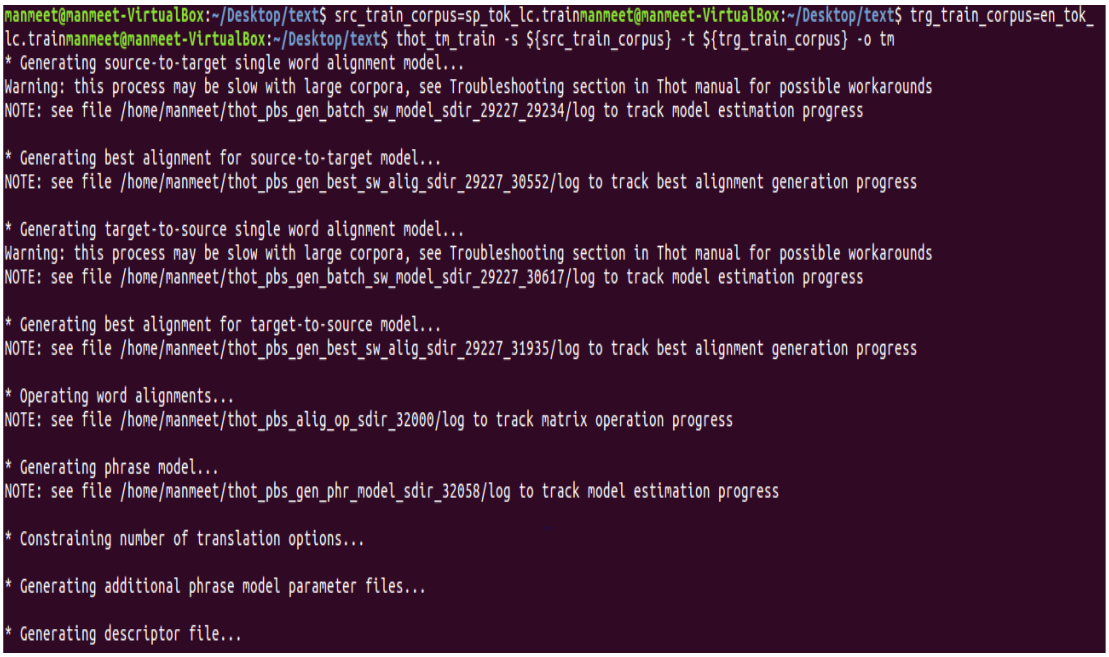


Figure 4.12 Translation Model Training

4.4.4. Basic configuration file generation

This provide us with the parameters utilised by Thot tool telling us the position of language and translation model, weights in log linear model etc. thot_gen_cfg_file tool is incorporated for this. This step is executed once we have trained language and training models. Figure 4.13 shows us the command for generating configuration file.

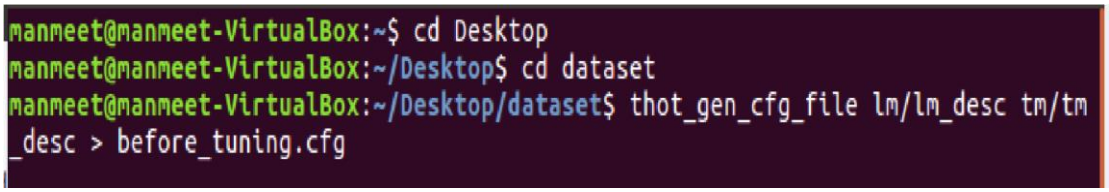


Figure 4.13 Configuration File Generation

4.4.5. Parameter tuning

Tuning phase effects the weights used in log linear models. Thot uses downhill-simplex algorithm for calculating weights and incorporates thot_smt_tune tool for this. Some basic input parameters are given in table 4.6 which include processors needed, source and target file, output directory and configuration file.

Table 4.6 Commands for Tuning

-pr(int)	Processors needed for estimation
-c(string)	Configuration file

-s(string)	Source sentences file
-t(string)	Target sentences file
-o(string)	Directory for output file

The figure 4.14 shows the commands for parameter tuning and depicts the output.

```
manmeet@manmeet-VirtualBox:~/Desktop/dataset$ src_dev_corpus=${PREFIX}/home/manm
eet/Desktop/dataset/sp_tok_lc.dev
manmeet@manmeet-VirtualBox:~/Desktop/dataset$ trg_dev_corpus=${PREFIX}/home/manm
eet/Desktop/dataset/en_tok_lc.dev
manmeet@manmeet-VirtualBox:~/Desktop/dataset$ thot_smt_tune -c before_tuning.cfg
 -s ${src_dev_corpus} -t ${trg_dev_corpus} -o smt_tune
* Tuning language model...
```

Figure 4.14 Parameter Tuning

4.4.6. Phrase based filtering

Phrase models have many parameters for large training corpus which make it impossible to store in main memory. Hence we filter out phrase parameters. Since our is small corpus this step is omitted. thot_prepare_sys_for_test tool is incorporated in Thot tool for providing this functionality.

4.4.7. Search

As already discussed branch and bound algorithm is used for search. We have thot_decoder tool for providing this functionality. The figure 4.15 shows the output of the system on terminal for a particular sentence "भवान् कथम् अस्ति ".given in Sanskrit.

```
time to create hyps  0.020(15%)
        estimate score  0.000(0%)
        calc lm         0.010(7%)
        manage stack    0.040(30%)
        other           0.060(46%)
total source words = 3
     words deleted = 0()
    words inserted = 0()
search took 0.130 seconds
Best Translation:आप कैसे है [1111]  [total = 0.877] <<0.000, -4.000, 0.000, -0.511,
 0.000, -3.429]
Best Hypothesis generation time : [477.000] seconds
sentence Decoding Time: [477.000] seconds
Finished translating

output (END)
```

Figure 4.15 Search Output

4.5 Cloud Deployment:

Cloud helps in providing flexible resources (software) on request. These resources are provisioned as services to the users according to their usage and are charged accordingly. Cloud is developing as an expansive field of research with the goal of investigating the immense measure of data and client solicitations to extract information. There are three types of cloud models: private, public and hybrid. Cloud models provide us with three services: Infrastructure as a service, platform as a service and software as a service.

4.5.1 Architecture of cloud deployment

Virtual Appliance can run on standalone Virtual Machine or available on cloud. A **virtual private server** (**VPS**) is a type of hosting service which run its own copy of operating system is used to host the proposed system on cloud. VPS is created by process of virtualization which helps to divide one physical server into multiple servers and then each server can run multiple applications and operating system. It provide a user with dedicated server. Virtual private server can be purchased from godaddy, bigrock and can be configured according to the requirements of CPU, memory and processors. Figure 4.16 shows the architecture for the system hosting on cloud using VPS:

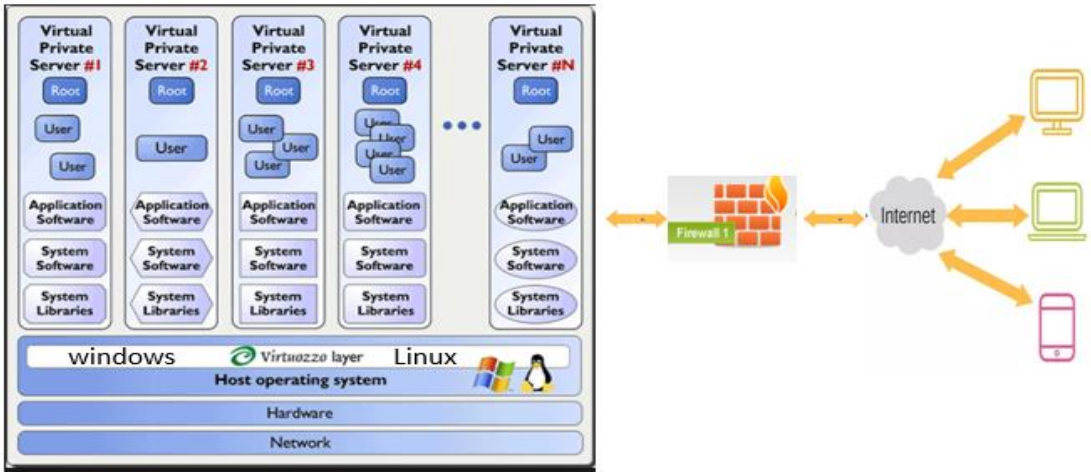


Figure 4.16 Architecture of VPS cloud hosting [40]

While creating server on Ubuntu 16.04 few configuration steps needed to be followed which are given below:

i. Root Login- for logging in to server provide IP address and password or can be installed by SSH by following command:

$ ssh root@ ip_address

ii. Create new user- since root is administrative user with high privileges it is not recommended to use root account for regular basis hence a new account is created with following command and creates new user called manmeet:

adduser manmeet

Then fill in with password and other details.

iii. Root previligies-Sometimes administrative privileges are needed by the new user therefore add new user to sudo group since sudo helps in running administrative commands

usermod –aG sudo manmeet

iv. Public key authentication- for securinf the server public key authentication is needed. To generate key pair command is:

$ ssh-keygen

Then obtain the output and provide with file name and path in will be used for saving the key. This will provide us with public and private key. Once the keys are created, copy the public key to new server. For doing this following steps need to be followed on the terminal of our computers:

$ cat ~/ .ssh/id_rsa.pub

This command will print the contents of the public key where id_rsa.pub refers to public key generated. Copy this to clipboard. Go on the server as root user enter command to switch the user and create directory called .ssh and restrict its permission to the authorised authorized_keys with following commands:

$ mkdir ~/.ssh

$ chmod 700 ~/.ssh

$ chmod 600 ~/.ssh/authorized_keys

v. Disable password authentication- to increase security of the server diable password only authentication. This will restrict access to SSH access for the server to public key authentication therefore login will be successful only if private key pair with public key.

$ sudo nano /etc/ssh/sshd_config

sshd_config- Disable password authentication

passwordAuthentication no

and reload SSH daemon by typing

$ sudo systemctl reload sshd

vi. Test log in –login to server using account created:

$ ssh manmeet@your_server_ip

vii. Set up firewall-these are used in Ubuntu 16.04 to make connections to certain services allowed. Open SSH to connect to server, with registered UFW (different applications can register their profiles with UFW)

$ sudo ufw app list

$ sudo ufw allow openSSH

$ sudo ufw enable

For installing any new software on server, create a new folder www called scripts and add test test.py to it. In order to run the translator system we need to add it server and make use of scripts to run the system. Script takes the input, compute them and returns the output. The commands mentioned in data processing are needed to run here in the following command:

$message = exec("/var/www/scripts/test.py 2>&1")

Print_r($message);

After running all the steps we get the system deployed on the server and hence it is seen as depicted in the figure 4.17

Figure 4.17 Results of Translation

4.6 Conclusion

This chapter shows the steps undertaken for implementing the proposed system. It presents different models used for the working of the system which describes n gram language model, phrase based alignment translation model and Branch and bond

algorithm for search used for building Statistical Machine Translation System. The system is later deployed on cloud. Steps undertaken for deploying the system on cloud using Virtual Private Server are mentioned briefly. Next chapter shows evaluation of the experimental results obtained by the proposed system and discusses them briefly.

Chapter 5
Experimental Results and Evaluation

5.1 Evaluation of the System

Evaluation of the proposed Sanskrit to Hindi SMT system (accept Sanskrit as source language and provide Hindi as target language) is tested by Manual and automated methods. Manual evaluation has been done on parameters of accuracy and fluency to evaluate the system. Automated evaluation with help of Microsoft Translation hub is used for calculating bilingual evaluation understudy (BLEU) score. Accuracy refers to the degree to which translation has been achieved with reference sentence. Fluency is measured as degree of grammatical accuracy attained in translation text. Evaluation done on different levels to calculate fluency and adequacy by two persons as depicted in table 5.1 and 5.2 respectively is shown below:

Table 5.1 Levels of Fluency

Parameter	Meaning	Rank according to Human evaluation
Perfect	Indicates good grammar of translated sentence	4
Fair	Translated sentence lack of grammar but easy to understand.	3
Acceptable	Broken but understandable translated sentence.	2
Bad	Translation not clear/understandable.	1

Table 5.2: Levels of Adequacy

Parameter	Meaning	Ranking
Most	Translated sentence conveys meaning of source sentence	3
Some	Translated sentence conveys some meaning	2
None	Translated sentence does not convey the meaning	1

After evaluating by two persons geometric average is calculate as depicted in table 5.3 below:

Table 5.3: Geometric Average of Adequacy and Fluency

	Adequacy	Fluency
Geometric average	2.99	2.63

Table 5.1 given above shows ranking given by human evaluation in terms of fluency showing that the sentences gives output with almost correct grammar. Table 5.2 shows ranking according to adequacy and it is cleared from the ranking above that the translated sentences preserve the meaning of the input text. Table 5.3 gives geometric average of the two parameters i.e. fluency and adequacy which is achieved to be 2.63 and 2.99 respectively.

Bleu refers to bilingual evaluation understudy score metric which helps in evaluation of generated sentence with reference sentence [31]. Bleu score is calculated with help of MT hub, a fully integrated system for developing translation system and is achieved to be 12.04. Translation of some sentences using the system developed has been shown in table 5.4 below:

Table 5.4: Translation from Sanskrit to Hindi

s.no	Sanskrit sentences	Hindi sentences
1	भवान् कथम् अस्ति ?	आप कैसे है?
2	भवतः नाम सकम	भव से नामक
3	श्रीमन् कृपया अत्र आगच्छतु	कृपा से यहाँ आये
4	सीता कदा पठसत	सीता कब पढती है
5	अहम् खादासम अत्र भ जनं	मैं खाता हाँ यहाँ भ जजन
6	वसन्तः रमणीयः ऋतुः अस्ति । इदानीं शीतकालस्य भीषणा शीतलता न भवसत । मन्दं मन्दं वायुः चलती । सवहंगाः कूजस्तन्त । सवसवधैः कुसुमैः वृक्षाः आच्छासदताः भवस्तन्त । कुसुमेषु भ्रमराः गुज्जस्तन्त । धान्येन धरणी पररपूणाा भवसत ।	वसन्त एक सुन्दर ऋतु है इस समय शीत काल की तरह भीषण ठंडा नहीं रहता है धीरे-धीरे हवा वहती है पनी गाते है सवसभन्न प्रकार के फूल ं से वृक्ष भर जाते हैं फूल ं पर भौरा गुंजते हैं पृथ्वी धान से भर जाता है
7	श्वः मम जन्मसदनम् अस्ति।	कल मेरा जन्मसदन है
8	अशिशित: मनुष्य: अङ्गुष्ठेन मुद्ा ं करोशत।	अनपढ व्यक्ति अंगूठे से शिह्न करता है

9	प्रेरकः सूचकश्चैव वाचक दशाकिथा । सशक्षक ब धकश्चैव षडेते गुरवः स्मृताः ॥	प्रेरणा देनेवाले, सूचन देनेवाले, सच बतानेवाले, रा ा सदखानेवाले, सशक्षा देनेवाले, और ब ध करानेवाले – ये सब गुरु समान है
10	भारतदे स्य राजधानी नवदेहली ।	भारत की राजधानी नई शदल्ली है

After translation of approx. 50 sentences we came with following conclusion:

i. Sometimes system generates less words in translated text as shown below: कृपया उपविशतु। (Sanskrit sentences)

बैठिए। (Hindi sentences)

ii. Accuracy

Accuracy refers to the quality of being precise and correct. As the number of word counts in the input sentence given to the system increases the accuracy of the output generated by the translation system decreases. In order to attain greater accuracy quantity and quality of the corpus should be increased. Table 5.5 shows the accuracy of the proposed system with respect to the word count.

Table 5.5 Word Count v/s Accuracy

Word Count	Accuracy
2	100
4	94
6	87
8	52
10	30

Graph in figure 5.1 represents word count v/s accuracy where a decrease in accuracy slope is shown with increase in the words in input sentences.

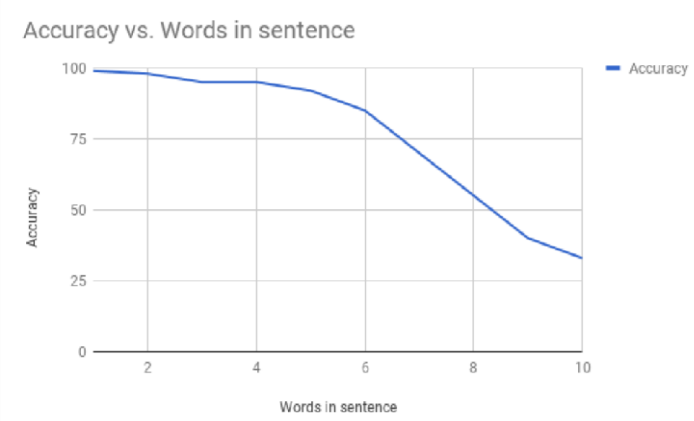


Figure 5.1 Graph for Word Count v/s Accuracy

iii. Reordering problem

The translated sentences sometimes gives error in verb translation by not ordering them in write way. Example is demonstrated below:

अहम् खादासम अत्र भ जनं (Sanskrit sentence)

मैं खाता हूँ यहाँ भोजन (Hindi Sentences)

iv. Error while ending the sentence

Error while ending of the translated text is observed in the proposed Sanskrit to Hindi translation system. The Translated Hindi text does not end with proper punctuation mark instead ends with numbering as shown in figure 5.2 below:

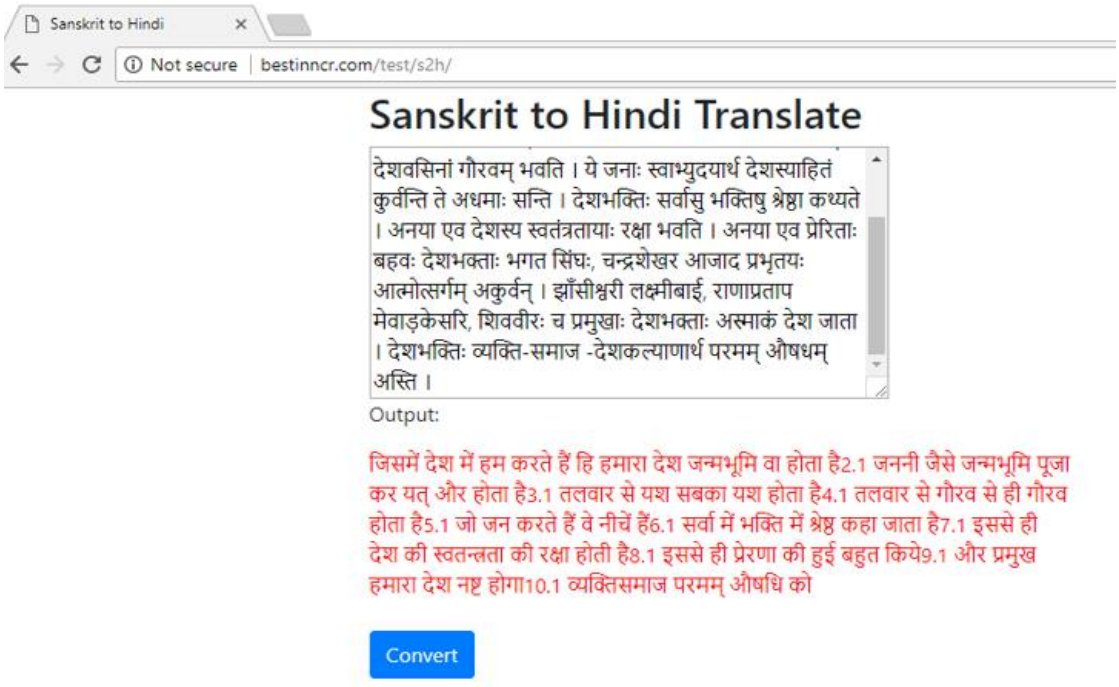


Figure 5.2 Translation for Paragraph

From the above observations it is concluded that for increasing the accuracy rate of the system, the dataset needs to be increased with more words, grammar, lexical and structure ambiguity.

5.2 Conclusion

After reviewing the translation system by human evaluation to check adequacy which is attained to be 2.99 and fluency 0f system with score of 2.63, Bleu score obtained with the help of MT hub translator i.e. 12.04 the chapter concludes that the translation system developed gives accurate results with sentences of 8-15 words and may suffer from some errors with paragraph translation. In order to get more accurate translations corpus size needs to be increased both in quality and quantity wise. The next chapter gives us overall conclusion of the thesis and gives future scope in the field of Machine translation.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

The research work undertaken represents Sanskrit to Hindi translation system using statistical based approach. This is corpus based methodology which uses parallel corpus of 1000 Sanskrit to Hindi sentences generated manually for training the system. The main problems of SMT i.e. availability of corpus, training and search have been taken into account and solved by making corpus of 1000 sentences manually, using n-gram language model, phrase alignment model for translations model and branch and bound algorithm for search respectively. The system takes Sanskrit as input and gives corresponding Hindi sentences as output. The accuracy of system is evaluated using human and automated evaluation method. Parameters like fluency, adequacy are calculated manually and Bleu score is evaluated using MT Translator. System is helpful in many areas since we have large amount of literature in Sanskrit. The quality of the translation system obtained rely upon quality and quantity of bilingual parallel corpus.

6.2 Future Scope

Future directions for the represented Sanskrit to Hindi SMT system are given below:

i. Inclusion of parallel corpus for multiple languages in source-target pair which will help in building of Multilingual Translation systems.

ii. Increasing corpus of the language-pair so as to increase translation quality since translation quality highly depends quantity and quality of the dataset used.

iii. Another advancement which can be done is developing a mobile application in which message containing Sanskrit text is converted into Hindi language.

iv. In order to improve the quality of translation, corpus can be preprocessed to change its clause structure.

v. In order to improve human evaluation score, the translated text obtained needs to be reordered and processed by post-processing to solve grammatical mistakes.

REFERENCES

[1] Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema RamírezSánchez, and Francis M. Tyers,"Apertium: a free/open-source platform for rule-based machine translation," *Machine translation*, vol.25, no. 2, pp. 127-144, 2011.

[2] Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal, "ANUSAARAKA: Machine Translation in stages", Vivek, a quarterly in Artificial Intelligence, NCST Mumbai, vol. 10, no. 3, pp. 22-25, 1997.

[3] Akshar Bharti, Chaitanya Vineet, Amba P. Kulkarni & Rajiv Sangal, "ANUSAARAKA: overcoming the language barrier in India", published in Anuvad: approaches to Translation, pp.456-468, 2001.

[4] G. S. Josan and G. S. Lehal, "A punjabi to hindi machine translation system", 22nd International Conference on Computational Linguistics: Demonstration Papers. Association for Computational Linguistics, 2008, pp. 157–160, 2008.

[5] Vishal Goyal & Gurpreet Singh Lehal, "Web Based Hindi to Punjabi Machine Translation System", International Journal of Emerging Technologies in Web Intelligence, ACADEMY PUBLISHER, Vol. 2, no. 2, pp. 148-151, 2010.

[6] Darbari, H, "Computer-assisted translation system–an Indian perspective", In Machine Translation Summit VII*",* (pp. 80-85), 1999.

[7] Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma & R. Sangal,"Machine Translation: The Shakti Approach", Pre-Conference Tutorial, ICON-2003. (pp. 478-485), 2003.

[8] Murthy. K, "MAT: A Machine Assisted Translation System", In Proceedings of Symposium on Translation Support System, IIT Kanpur, pp.134-139, 2002.

[9] Mumbai, C. D. A. C,"MaTra: an English to Hindi Machine Translation System", a report by CDAC Mumbai formerly NCST , pp. 211-216, 2008.

[10] Sampark: Machine Translation System among Indian languages http://tdildc.in/index.php?option=com_vertical&parentid=74, http://sampark.iiit.ac.in/ , 2009.

[11] Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R. & Jain, A.,"ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages," IEEE International Conference on: Systems, Man and Cybernetics, Intelligent Systems for the 21st Century, pp.1609-1614, 1995.

[12] Sinha, R.M.K. and Jain, A., "AnglaHindi: an English to Hindi machine-aided translation system," MT Summit IX, New Orleans, USA, pp.494-497, 2003.

[13] Dave, S., Parikh, J., & Bhattacharya, P, "Interlingua-based English-Hindi Machine Translation and Language Divergence", Journal of Machine Translation, vol. 34, no. 4, pp. 251-304, 2001.

[14] Antony, P.J., "Machine translation approaches and survey for Indian languages", International Journal of Computational Linguistics & Chinese Language Processing, vol. 18, no. 1, pp. 123-128, 2013.

[15] Lopez A., "Statistical machine translation", ACM Computing Surveys (CSUR), vol.40, no. 3, pp. 567-578, 2008.

[16] Koehn P, Och J, Daniel Marcu,"Statistical Phrase-Based Translation", Proceedings of HLT, NAACL,Edmonton, Main Papers, pp.485, 2003.

[17] S.K. Dwivedi and P. P. Sukadeve, "Machine Translation System Indian Perspectives", Proceeding of Journal of Computer Science vol. 6 no. 10, pp 1082-1087, 2010.

[18] Chiang, D, "Hierarchical phrase-based translation", computational linguistics, vol.33, no. 2, pp. 201-228, 2007.

[19] Rahul C, Dinunath K ,RemyaRavivardhan ,K.P Soman ,"Rule Based Reordering and Morphological Processing for English Malayalam Statistical Machine Translation", Advances in Computing, Control, & Telecommunication Technologies,. ACT,pp. 345-351, 2009.

[20] Ahsan A., Kolachina P., et. al., "Coupling Statistical Machine Translation with Rule-based Transfer and Generation", AMTA- The Ninth Conference of the

Association for Machine Translation in the Americas. Denver, Colorado, pp. 234-242, 2010.

[21] Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun et al., "Google's neural machine translation system: Bridging the gap between human and machine translation", arXiv preprint arXiv:1609.08144, Pp.201-223, 2016.

[22] Nalluri, Anitha, and Vijayanand Kommaluri, "Statistical Machine Translation using Joshua: An approach to build “enTel” system", Parsing in Indian Languages, pp. 465-474, 2011.

[23] Sindhu, D. V., and B. M. Sagar, "Study on machine translation approaches for Indian languages and their challenges", In Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference on, pp. 262-267. IEEE, 2016.

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

[25] Vijayanand, Kommaluri, Sirajul Islam Choudhury, and Pranab Ratna, "Vaasaanubaada: automatic machine translation of bilingual bengali-assamese news texts", In Language Engineering Conference, 2002. Proceedings, pp. 183-188. IEEE, 2002.

[26] Ambati, V &Rohini, U, “A Hybrid Approach to EBMT for Indian Languages”, ICON, Pp.567-576. 2007.

[27] Kituku, Benson, Lawrence Muchemi, and Wanjiku Nganga, "A Review on Machine Translation Approaches", Indonesian Journal of Electrical Engineering and Computer Science vol. 1, no. 1, pp. 182-190, 2016.

[28] Behera, Pitambar, Renu Singh, and Girish Nath Jha.,"Evaluation of Anuvadaksh (EILMT) English-Odia Machine-assisted Translation Tool", WILDRE: LREC , vol. 56, no. 6, pp. 567-573, 2016.

[29] Garje, G. V., and G. K. Kharate, "Survey of machine translation systems in India", International Journal on Natural Language Computing (IJNLC) vol 2, pp.47-67, 2013.

[30] Chatterji, S., Sonare, P., Sarkar, S., & Basu, A., "Lattice Based Lexical Transfer in Bengali Hindi Machine Translation Framework", In Proceedings of ICON: 9th International Conference on Natural Language Processing pp. 784-790, 2011.

[31] Papineni, K., Roukos, S.Ward, T., & Zhu, W. J., "BLEU: a method for automatic evaluation of machine translation", In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318, 2002.

[32] B. P. Rimal and E. Choi, "A taxonomy and survey of cloud computing systems", in Fifth International Joint Conference on INC, IMS and IDC, Seoul, Korea, vol. 67, no. 2, pp. 678-680, 2009.

[33] Kituku, B., Muchemi, L. and Nganga, W., "A Review on Machine Translation Approaches", Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, no. 1, pp.182-190, 2016.

[34] Arnold D, Balkan L, Meijer S, Humphreys R L, Sadler L, "Machine Translation: an Introductory Guide", NCC Blackwell, London, pp. 78-89,1994.

[35] Siddiqui T, Tiwary U S, "Natural language processing and information retrieval" Oxford University Press, New Delhi, vol. 78, no. 6, pp. 456-460, 2008.

[36] Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan et al., "Moses: Open source toolkit for statistical machine translation", In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, vol.56, pp. 177-180, 2007.

[37] Saini, Sandeep, and Vineet Sahula. "A survey of machine translation techniques and systems for indian languages", In Computational Intelligence & Communication Technology (CICT), IEEE International Conference on, pp. 676-681, IEEE, 2015.

[38] Daniel Ortiz-Mart´ınez, Ismael Garc´ıa-Varea, Francisco Casacuberta, "Thot: a toolkit to train phrase-based models for statistical machine translation", In Proc. of the Tenth Machine Translation Summit (MT-Summit), Phuket, Thailand,pp 160-169, 2005.

[39] Daniel Ortiz-Mart´ınez, Francisco Casacuberta,“The New Thot Toolkit for Fully Automatic and Interactive Statistical Machine Translation”, In Proc. of the 14th Annual Meeting of the European Association for Computational Linguistics (ACL): System Demonstrations, pp. 45–48, April 2014.

[40] Sotomayor, Borja, Rubén S. Montero, Ignacio M. Llorente, and Ian Foster, “Virtual infrastructure management in private and hybrid clouds”, IEEE Internet computing, vol.13, no. 5, pp.201-210, 2009.

[41] Daniel Ortiz-Mart´ınez, “Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation”, PhD Thesis. Universitat Polit`ecnica de Val`encia. Advisors: Ismael Garc´ıa Varea and Francisco Casacuberta. 2011.

List of Publications

[1] Manmeet kaur, Inderveer Chana, Ravinder Kumar "Machine Translation Approaches for Indian Languages: A Survey", IEEE 4th International Conference for Convergence in Technology (I2CT) SDMIT MANGALORE. [ACCEPETED]

Manmeet_plag_report

ORIGINALITY REPORT

10%	8%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	daormar.github.io Internet Source	1%
2	www.ijert.org Internet Source	1%
3	dspace.thapar.edu:8080 Internet Source	<1%
4	airccse.org Internet Source	<1%
5	www.mt-archive.info Internet Source	<1%
6	Benson Kituku, Lawrence Muchemi, Wanjiku Nganga. "A Review on Machine Translation Approaches", Indonesian Journal of Electrical Engineering and Computer Science, 2016 Publication	<1%
7	dyuthi.cusat.ac.in Internet Source	<1%
8	D. V. Sindhu, B. M. Sagar. "Study on machine translation approaches for Indian languages	<1%