A NOVEL APPROACH FOR SENSE BASED HINDI TO TAMIL MACHINE TRANSLATION SYSTEM

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

K. VIMAL KUMAR
ENROLLMENT NO.
10403023

Department of Computer Science Engineering and Information Technology

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY
(Declared Deemed to be University U/S 3 of UGC Act)
A-10, SECTOR-62, NOIDA, INDIA

JUNE 2019

TABLE OF CONTENTS

CHAPTER 2

CHAPTER 3

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled "**A Novel Approach for Sense Based Hindi to Tamil Machine Translation System**" submitted at **Jaypee Institute of Information Technology, Noida, India**, is an authentic record of my work carried out under the supervision of **(Late) Prof. Padam Kumar**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. Thesis.

(Signature of Scholar)

K. Vimal Kumar

Department of Computer Science Engineering and Information Technology,

Jaypee Institute of Information Technology, Noida, India

Date: June, 2019

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled "**A Novel Approach for Sense Based Hindi to Tamil Machine Translation System**", submitted by **K. Vimal Kumar** at **Jaypee Institute of Information Technology, Noida, India**, is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

(Supervisor)	(Signature of Administrative Supervisor)
(Late) Prof. Padam Kumar	Prof. Vikas Saxena
Dean (R, I & D),	HOD, Dept. of CSE&IT,
Jaypee Institute of Information Technology, Noida, India.	Jaypee Institute of Information Technology, Noida, India.
Date: June, 2019	Date: June, 2019

ACKNOWLEDGEMENT

Completion of my doctoral thesis was possible with the support of several people. I would like to express my sincere gratitude to all of them. First, I would like to express my sincere gratitude for the time, encouragement, and guidance provided by my supervisor, (Late) Prof. Padam Kumar and whose help cannot be truly expressed in words. Indeed the technical insight and intuition have helped me in exploring in the area of machine translation system. I also would like to thank him for the guidance, valuable advice, motivation, enthusiasm, and immense knowledge that has greatly improved my dedication to carry on with research activities.

I would also like to thank Prof. S.C. Saxena, (Late) Prof. Padam Kumar, Prof. Sanjay Goel, and Prof. Vikas Saxena, Dr. Sandeep Kumar Singh and Prof. Alka Tripathi for serving on my final Ph.D committee and for all of their support and suggestions. Few faculty members of the JIIT, Noida have been very kind enough to extend their help at various phases of this research work, whenever I approached them, and I do hereby acknowledge to them. I would like to thank Prof. Bhu Dev Sharma, for the suggestion provided to me in sense based machine translation. I would like to thank Dr. Hema N, CSE Department for the support provided in writing. I would like to thank Prof. Hima Bindu Maringanti (ex-supervisor), Dr. Divakar Yadav (ex-supervisor) and Dr. Yamuna Prasad (ex-supervisor) for the constant support in reviewing, writing and suggestion provided in my research work.

I would also like to thank entire JIIT family members with whom I made during the past seven years on my path toward earning my doctoral degree. Thanks for all the great times and all the encouragement and suggestions. It made this journey much more pleasant and meaningful.

I owe a lot to my dear parents, who encouraged and helped me at every stage of my personal and academic life, and longed to see this achievement come true. My father, Prof K Kumar, being an active researcher in corrosion materials by profession, has developed the cement polymer composite coating (CPCC) and still working on various research works. I should thank him for being a motivation factor for this research work. He has been my advisor and great support from his field experiences.

Special thanks to my dear family. Simple words cannot express how grateful I am to my daughter Harshitha (3 years) for all the sacrifices that she had made on behalf of me, by waiting patiently for completion of my seven-year research work. Your love and prayer for me were what sustained me thus far. I would like to thank my mother, C. Vijayalakshmi, for providing me the support during hard times and also for praying on behalf of me for my research work. At the end I would like to express appreciation to my beloved wife G. Viduthalai Selvi who spent more time in handling my daughter to provide me the valueable time for my PhD and was always my support in the moments when there was no one to answer my queries.

Above all, I owe it all to the almighty God for granting me the wisdom, health and strength to undertake this research task and enabling me to its completion.

Thank you all.

ABSTRACT

Machine translation is the process of generating the target language text from the source language text without distorting the information being conveyed in the source language text. The machine translation system involves two different languages. In order to perform the translation there is need for mapping between these two languages which is a complex task. These mapping can be of various forms such as one-to-one mapping, one-to-many mapping, many-to-one mapping and many-to-many mapping. The mapping totally depends on the languages being considered for the translation. To have a more accurate translation, there is need for a mapping between the languages that uses the syntactic and semantic information. In most of the cases, the syntactic information is being considered for the mapping. To have a meaningful translation, there is need for semantic information-based mapping along with syntactic mapping. The aim of this research is to develop a Hindi to Tamil machine translation system using both syntactic and semantic information-based mapping. Hindi and Tamil language are basically free word order languages. Thus, it increases the complexity of mapping further. But the mapping that uses syntactic and semantic information will be helpful in translation between these two languages.

A statistical machine translation system was proposed which considers the syntactic and semantic information. The syntactic information is captured using the part-of-speech (POS) tagger over the source text, Hindi. The semantic information in the source text is extracted using latent semantic analysis (LSA) and Hindi wordnet. Both syntactic and semantic information is used in the statistical machine translation system. The statistical data required for translation are developed using the mapping between the languages by considering the syntactic and semantic information. The bilingual evaluation understudy (BLEU) [1] score is found to be 0.68 in the scale of 0 to 1. Since, Hindi and Tamil are low resource languages, there is need for an intermediate bridging language to improve the translation further.

Using a pivot language, a Hindi to Tamil statistical machine translation system was proposed. English language is chosen as the pivot language. The mapping between Hindi and Tamil is performed indirectly using English. The statistical data are developed by mapping Hindi words with English words, later by mapping English words with Tamil words. Initially, the semantic information was not considered in this pivot language-based translation. The BLEU score of the pivot-based machine translation is found to be 0.7394. Since Hindi and Tamil languages are rich in morphology as compared to English language, there is loss of semantics during the pivot-based translation. To retain the semantics that were distorted during this pivot-based translation the semantic analysis was introduced before the translation process and the BLEU score was found to increase slightly up to 0.7637. But, still there is need for improvement in the machine translation from Hindi to Tamil.

A neural machine translation was proposed using sequence to sequence model. The sequence to sequence approach makes use of long-short term memory (LSTM) neural network to map the languages. To capture the semantic and syntactic information, word embedding was performed in each of the languages using a continuous bag-of-words (CBOW) model. The BLEU score of the neural machine translation is found to be 0.7588.

LIST OF ABBREVIATIONS

BLEU	Bilingual Evaluation Understudy
CBOW	Continuous bag-of-words model
CBR	Case-based reasoning
DeitY	Department of Electronics & Information Technology
DIT	Department of Information Technology
EBMT	Example-based machine translation
E-ILMT	English to Indian Language Machine Translation
GIZA++	A word alignment tool
HMM	Hidden Markov Model
IBM	International Business Machine Corporation
ILCI	Indian Language Corpora Initiative
IT	Information Technology
LSA	Latent Semantic Analysis
LSTM	Long-Short term memory neural network
MC&IT	Ministry of Communication & Information Technology
MLP	Multilayer perceptron

MOSES	A Statistical machine translation tool
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural machine translation
POS	Part-of-Speech
RBMT	Rule-based machine translation
RNN	Recurrent Neural Network
SL	Source Language
SMT	Statistical machine translation
SOV	Subject Object Verb
SVO	Subject Verb Object
TAG	Tree Adjoining Grammar
TAM	Tense Aspect and Modality
TDIL	Technology Development for Indian Languages programme
TL	Target Language
UNL	Universal Networking Language
WSD	Word Sense Disambiguation

LIST OF SYMBOLS

pos	Part-of-speech of the current word
W_s	Source language word
W_t	Target language word
W_p	Pivot language word
L	Term matrix or Left singular matrix
R	Document matrix or Right singular matrix
S	Term by Document matrix or Singular diagonal matrix
f_{tn}^{m}	Frequency of n^{th}-term in m^{th} document
tag_s	Part-of-speech of source word
T_n	Part-of-speech of word at n^{th} position
T_{n-1}	Part-of-speech of word at $(n-1)^{th}$ position
W_n	Word at n^{th} position
x_i	i^{th} vector in right singular matrix
y_i	i^{th} vector in singular diagonal matrix
i	Word position in source language
j	Word position in target language

l	Length of sentence in source language
m	Length of sentence in target language
W_{prev}	Previous word in target language
$word_i$	Word at ith position
tag_i	Part-of-speech of word at ith position
tag_{i-1}	Part-of-speech of word at (i-1)th position
tag_j	Part-of-speech at jth index in the tagset
N	Number of distinct part-of-speech tags
W_j	Weight assigned to jth neuron
W_{p-1}	Previous word in the pivot language
W_{t-1}	Previous word in the target language
A	Term-document frequency matrix
W_{IH}	Weight matrix between Input layer and hidden layer
W_{HO}	Weight matrix between hidden layer and output layer
IN_H	Hidden layer input
O_H	Output of hidden layer
O_j	Predicted output vector of jth word
u_k	kth value of the vector ‘u’
N	Vocabulary Size

O_{pred}	Predicted output vector
O_{act}	Actual output vector
T_m	Target word at mth position
T_{m-1}	Target word at (m-1)th position
h_t	Output of hidden state at time ‘t’
o_t	Out gate value at time ‘t’
c_t	LSTM Cell value at time ‘t’
c_{t-1}	LSTM Cell value at time ‘t-1’
f_t	Forget gate value at time ‘t’
g_t	New cell value at time ‘t’
i_t	Input gate value at time ‘t’
x_t	Input vector at time ‘t’
h_{t-1}	Output of hidden state at time ‘t-1’
w_i^o	Weight vector between input layer and out gate
w_h^o	Weight vector between hidden layer and out gate
b_i^o	Bias value in between input layer and out gate
b_h^o	Bias value in between hidden layer and out gate
w_i^g	Weight vector between input layer and cell gate
w_h^g	Weight vector between hidden layer and cell gate

w_i^f	Weight vector between input layer and forget gate
w_h^f	Weight vector between hidden layer and forget gate
w_i^i	Weight vector between input layer and input gate
w_h^i	Weight vector between hidden layer and input gate
b_i^g	Bias value in between input layer and cell gate
b_h^g	Bias value in between hidden layer and cell gate
b_i^f	Bias value in between input layer and forget gate
b_h^f	Bias value in between hidden layer and forget gate
b_i^i	Bias value in between input layer and input gate
b_h^i	Bias value in between hidden layer and input gate

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In the current internet world, every information of interest is available in the huge web for the global market to sustain its business. However, the available information may not be in the native language of the end user. To understand the information provided in some other language, there is need for a translator that can provide the information in the end user's native language. Human translators were used in the olden days. Since 1940s automatic machine translation developed [2] which is defined as the process of extracting features of source language to generate the target language text by using these extracted features. Machine translation being a subarea of natural language processing, needs to learn the mappings in each of the languages being considered and map the words in source language with their corresponding words in target language. Based on these mappings, the machine translation should be able to translate the text from source to target language. But, these mapping of words is not always one-to-one, thus, increasing the complexity of machine translation system.

The mapping of words is categorized as – one-to-one, one-to-many, many-to-one and many-to-many. One-to-one mapping basically matches the single source word with the single target word. One-to-many maps the single word in source language to multiple words in target language. Many-to-one mapping is used for mapping the multiple words in source language to a single word in target language. In case of many-to-many, the mapping is between multiple words in source language and multiple words in target language. In some languages, this mapping is even more complex due to the occurrence of multiple target language words in different positions for a given source word. For an example on many-to-one mapping, consider the Hindi sentence in the first sentence pair given in Table-1.1, the words "***जा रहा है। (ja raha hai)***". These words will be mapped to a single target word "***போகிறார். (pōkiṟār)***" in Tamil. In Table-1.1, the second sentence pair has many-to-one as well as many-to-many kind of mapping. The many-to-many mapping is highlighted as bold and underlined text whereas, many-to-one is shown as bold/underlined. The one-to-one mapping is directly shown in respective column without any highlighting.

Table 1.1: Types of mappings between Hindi and Tamil language

S. No.	Language	Sentences			
1	English	Ravi **is going** to play.			
	Hindi	रवि (Ravi)	खेलने (Khelane)		**जा रहा है। (is going)** (ja raha hai.)
	Tamil	ரவி (Ravi)	விளையாட (viḷaiyāṭa)		**போகிறார். (is going)** (pōkiṟār)
2	English	I'm **going to playground** for playing.			
	Hindi	मैं (I'm) (main)	खेलने के लिए (khelane ke lie)	**खेल के मैदान में** (khel ke maidaan mein)	**जा रहा हूँ। (going)** (ja raha hoon.)
	Tamil	நான் (I'm) (Nāṉ)	விளையாடுவதற்கு (viḷaiyāṭuvataṟku)	**விளையாட்டு மைதானத்திற்கு** (viḷaiyāṭṭu maitāṉattiṟku)	**போகிறேன். (going)** (pokiṟēṉ.)

In languages which are morphologically rich like Hindi and Tamil, the probable translation for a word in source language can be different according to word's morphology. The Hindi and Tamil words embed the tense, aspect and modality (TAM) information with them. The words in target language should retain the grammatical information that is being used in the source word. Few examples of morphological richness of Hindi and Tamil as compared to English language is shown in Table-1.2.

Table 1.2: Analysis of word morphological structure in Hindi and Tamil

S. No.	Hindi Word	English Word	Tamil Word	Grammatical information
1	खा (kha)	Eat	சாப்பிட (Cāppiṭa)	Verb
2	खाएँ (khaen)	Eat	சாப்பிடுங்கள் (Cāppiṭuṅkaḷ)	Plural verb
3	खाउंगी (khaungee)	Will eat	சாப்பிடுவாள் (Cāppiṭuvāḷ)	Singular, Feminine Verb, First person
4	खाऊंगा (khaoonga)	Will eat	சாப்பிடுவேன் (cāppiṭuvēṉ)	Singular, Masculine Verb, First person
5	खाएगा (khaega)	Will eat	சாப்பிடுவீர்கள் (Cāppiṭuvīrkaḷ)	Singular, Masculine Verb, Second/third person
6	खाएगी (khaegee)	Will eat	சாப்பிடுவீர்கள் (Cāppiṭuvīrkaḷ)	Singular, Feminine Verb, Second/third person

The source and target languages being considered can be originated from same parent language or different. If these languages are originated from same parent language, then the machine translation approach that is being used for a language pair can be used for any other language pair that originated from the same parent language. For example, English and French – both these languages are originated from Latin and are also called as fixed word order languages. In a fixed word order language, the grammar is rigid and if the words are reordered then, the grammar as well as the meaning of sentence is lost. The machine translation mechanism being used for translating English to other language can be extended and used for French. But if the languages are totally unrelated based on the language specific features due to different origin, then a new model will be required to perform the translation. For example, English and Hindi – both these languages differ a lot based on the features. Thus, an approach used for translating English language to some other may not give fruitful result when applied on Hindi language. The existing machine translation approaches that can be used with required modifications according to the languages considered are,

- Rule based machine translation
- Statistical machine translation
- Transfer based machine translation
- Interlingua
- Example based machine translation
- Hybrid machine translation
- Neural machine translation

Each of these approaches can be applied between any language pairs. The rule-based machine translation system basically generates a ruleset for mapping the words from source to target. In general, a machine learning algorithm is applied to extract the ruleset from the parallel corpus. But since natural language is dynamic in nature, the rule-based system's performance degrades since the ruleset won't be mature enough to handle all such cases. The statistical machine translation system basically applies statistical model over the bilingual corpus and predicts the probable translation for the given input text. The statistical approach has a drawback of not being able to translate the text not fed during the training phase of the model. And also, out-of-vocabulary words will lead to poor translation accuracy.

The interlingua approach for machine translation makes use of an intermediate representation between the source and target language. This intermediate representation should be a representation which does not have dependency with any of the source and target language. It should also capture the semantics of the text being used. The demerit of this interlingua approach is its complexity in extracting the meaning from text and encode it on the intermediate representation so that it can be used while generating the target text [3]. The transfer-based approach works in similar way as interlingua, but the intermediate representation being used has some partial relationship with the source or target text. This is just to ease the way to capture the semantics and construct the intermediate representation. But, still decoding the semantics from the intermediate representation is a big challenge in the transfer-based approach.

In an example-based machine translation, the translation is based on the example bilingual corpus and it basically uses analogy. The translation works fine for sentences pairs that matches with the example sentences stored in the corpus. It is computationally complex due to the construction of dependency tree during the analysis and generation phases. A hybrid approach can be developed using any combination of the other approaches. A hybrid system developed with high efficiency is designed with a combination of statistical and rule-based approach. These approaches have their own merits and demerits according to the language pair being considered. To develop hybrid machine translation system using these approaches, there is need for a tradeoff between the merits and demerits of the approaches such that it provides an improvement on the accuracy of the system. This is due to the language specific features such as morphology of the words, part-of-speech tags, word order etc. Thus, to choose an approach for the translation process, there is need for analysis on various features of the languages. Perhaps each language pair considered for machine translation has its own difficulties and challenges.

1.2 MOTIVATION

The machine translation system has a wide use in the area of multi-lingual information retrieval, cross lingual summarization and so on. These applications have further variants such as text to text translation, text to speech translation, speech to text translation and speech to speech translation. In text to text translation, a given source text is translated to its corresponding target text. In case of text to speech translation, a given source text is translated to target language and is provided to the end user in audio form. If an end user speaks in his native language, the speech to text system analyzes the speech of the user to translate in target language. In case of speech to speech system, the user speaks in the native language and the speech is analyzed to convert into speech in the target language. This research is focused on text to text translation system from Hindi to Tamil. Hindi is a major Indian language used in almost whole of northern India (more than 50% of Indian population), whereas Tamil is majorly spoken and used in southern part of India by about 6% of population. Based on the analysis of these languages and the development of various NLP tools on the languages under consideration, the following are the prominent issues that motivated to propose a solution which can handle it in an appropriate manner,

i. Existing machine translation system for Indian language pairs needs improvement on the accuracy by a reasonable amount
ii. Word alignment model on these two languages needs to be modified
iii. The existing systems do not consider the contextual information of the words during machine translation
iv. In general, the resources for Indian languages are limited.
v. As compared with English, Hindi and Tamil are free word order languages.
vi. Semantics are lost during translation from Hindi to Tamil using English as a pivot.

There are many existing machine translation systems on Indian language pairs, but there is still need for considerable improvement on the accuracy of these systems. The languages such as Hindi and Tamil need a hybrid system to handle their language specific features such as free word order, morphological richness etc. The free word order feature in these languages gives the flexibility in reordering of words in a sentence without any loss of syntactic and semantic features. Morphologically rich feature provides more detail during the word formation. So, the words have the information such as tense, aspect and modality (TAM).

Since languages under consideration are morphologically rich and free word ordered, the rule-based approach is not an appropriate option to develop a machine translation system for these languages. The rule-based approach will give poor accuracy if the sentence has been reordered.

The statistical machine translation approach needs word alignment between source and target text. In case of Hindi and Tamil language, the alignment models being used for English language are not appropriate for these languages. This is due to the morphological richness of both the languages. The existing alignment models need to consider additional information from the corpus to perform a proper alignment between the source and target texts. Any of the existing approaches will be appropriate for translation from Hindi to Tamil but, a hybrid approach will be more suitable for these two languages.

Also, the translation performed by the existing systems does not consider the semantic features of the languages being used. Thus, it generates a target sentence which does not convey the actual meaning of the words mentioned in the source text. To make the translation without loss of semantics in the source text, there is need for word sense-based machine translation approach. To develop such a system there is need for language resources, but unfortunately, the resource available on the languages under consideration is not sufficient. Thus, an intermediate bridging language will be helpful to handle such low resource language issues. This intermediate language should basically have vast amount of resources to aid in the translation from source to target. This intermediate language can be either related to anyone of the languages being considered or not related to any of them. But, the intermediate language may affect the performance of the overall system due to corpus quality. To handle such scenario, there is need for deep learning approach also to improve the overall accuracy of the system without loss of syntactic and semantics.

1.3 CHALLENGES

There are major identified motivations in the development of Hindi to Tamil machine translation system and the motivations are – context-based machine translation, handling low resource availability and improve the accuracy of translation. To handle these motivating factors, there is a need to design a machine translation system which can address the following issues,

i. How to handle the low resource availability issue?
ii. How to perform word alignment between Hindi and Tamil language?
iii. How to extract the contextual information and incorporate it during the translation mechanism?

iv. How to retain the syntactic structure of the source text for further use in the following phases?

v. How to generate the target text in a grammatically correct manner?

vi. To improve the overall accuracy of the machine translation system

vii. Does the word embedding of Hindi and Tamil language capture the semantics?

The system designed in this research handles the low resource availability issue with the help of an intermediate natural language, called as pivot language. Thus, dividing the overall task of translation from source to target into two subtasks – source to pivot language and then pivot language to target. These subtasks can be handled by using any of the approaches briefed in section-1.1 and selection of approach is based on the language pairs being used. In general, the pivot-based machine translation system makes use of statistical approach. The pivot language assists in the translation process from source to target. The choice of the pivot language has to be in such a way that it can assist in the translation process and it has abundant resources.

In Hindi and Tamil language due to their morphological richness, there is need for modification on the word alignment algorithm. This is due to existence of multiple auxiliary verbs supporting the main verb in Hindi. But, in case of Tamil language, the auxiliary verb is suffixed along with the main verb. During translation from Hindi to Tamil, these multiple auxiliary verbs in Hindi have to be translated to a single word in Tamil. Thus, this increases the need to modify the existing alignment models. In this research, word alignment is performed by considering part-of-speech of the words so that the alignment is proper with respect to Hindi and Tamil languages.

Further in Hindi and Tamil, there are semantically equivalent words which are used in different contexts. For example, the words "*Kal"* and *"Aam"* in Hindi have different senses according to the sentence where it is being used. *"Kal"* can mean both yesterday or tomorrow, similarly, *"Aam"* can mean common man or the fruit mango depending on context. There are different words which have same meaning and have restriction in the usage according to the sentence. For example, *"Makaan (house)"* and *"Ghar (home)"* in Hindi have same sense and in Tamil both these words will be mapped to *"Veedu"*. Thus, there is need for contextual information during machine translation and to extract the contextual information, a word sense disambiguation phase needs to be introduced before the transfer phase. This word sense disambiguation phase extracts the contextual information and helps in finding the probable target word for the input source word. After the introduction of pivot language in between the source and target text, there is more semantic distortion happening during the translation. This semantic loss is due to the use of three different languages in translation. This can be taken care of by introducing sense disambiguation phase to extract the semantic features. Thus, a pivot-based system that considers a word sense is of greater need. Still the accuracy of the system has scope for further improvement.

Both Hindi and Tamil languages have words whose semantics changes according to its usage in the sentence. The change may be due to the change in part-of-speech of the word in the sentence. For example, the word *"Kar"* can be used as noun as well as verb. The word *"Taaza"* in Hindi can be used as adjective and also as an adverb. In both the cases, the word meaning changes according to its part-of-speech. Thus, the impact of syntactic information in machine translation from Hindi to Tamil is significant. The syntactic information also contributes to the proposed machine translation from Hindi to Tamil. During the translation to Tamil, the syntactic information of Hindi has to be considered.

Once the translation is generated in target text, there is need for grammatical correctness in the generated text. Since, the grammar of both languages may differ, there is need for rearrangement of words in a grammatically correct manner in accordance to the grammar of target text. In Hindi and Tamil language, there is use of subject-object-verb form. But, there is need for rearrangement phase in this as well. In the proposed system, there is a rearrangement phase after the transfer phase. In case of neural machine translation, this is performed by using attention mechanism.

The introduction of pivot language in between Hindi and Tamil machine translation was providing a considerable improvement but, after a certain threshold, the accuracy of the system gets saturated. Thus, to improve the accuracy further, there is need for more resources. This gives rise to a necessity for the neural machine translation. A neural machine translation system learns the features of both the language using the word vectors.

It has been mentioned above about the importance of syntactic and semantics in machine translation from Hindi to Tamil. The neural machine translation system should be trained to learn both these features from the training corpus. Thus, a word embedding is performed to capture the syntactic and semantic feature of both the languages. The word embedding phase that is being introduced before training the neural machine translation system and is used to generate the word vectors.

1.4 THESIS OBJECTIVE

The thesis presented here has four main objectives that are considered while designing a word sense-based Hindi to Tamil machine translation system. The solution approach for these objectives are discussed over here and in the next section of this chapter.

Objective-1: To explore the Hindi to Tamil statistical machine translation system that can perform a sensible translation

The existing Hindi to Tamil statistical machine translation system focuses more on the syntactic feature than semantic features. There are words in Hindi which are mapped with multiple different words in Tamil as shown in Table 1.3. But, this mapping can be narrowed down to some set of word mapping by using the part-of-speech of the word. Further, it can be reduced still to a smaller set of words with the help of contextual information. The first example word mentioned in table below is showing a single word in Hindi having equivalent Tamil word that has the same meaning. In case of the example-2 in Table-1.3, the same Hindi word has different senses according to its part-of-speech. Thus, this kind of examples will require syntactic analysis for choosing the proper Tamil word. The last three examples given in Table-1.3. has Hindi word having different sense according to the sentence in which they are being used whereas in Tamil there are separate words for each meaning. Thus, the impact of semantics and syntactic information is more on machine translation from Hindi to Tamil. The syntactic information is extracted by using part-of-speech tagger and the contextual information is extracted with the help of the word sense disambiguation module.

Table 1.3: One to Many mappings from Hindi to Tamil language

S. No.	Hindi Word	English Equivalents	Tamil Words
1	खाना (Kana)	Food	சாப்பிட (Sappita), உணவு (Unavu)
2	कर (Kar)	Do, Tax	வரி (Vari), செய்ய (Ceyya)
3	ताजा (Taaza)	Fresh (used for food/drinks), New (used for newspapers)	புதிய (Putiya), சுத்தமான (Cuttamāṉa), புத்துணர்ச்சி (Puthunarchi)
4	कल (Kal)	Yesterday, Tomorrow	நாளை (Nāḷai), நேற்று (Nētṟu)
5	आम (Aam)	Common, Mango	பொது (Potu), மாங்கா (Māṅkā)

Objective-2: To handle the low resource availability issue

The machine translation system designed to provide a sensible translation needs a large set of resources such as parallel corpus, NLP tools etc. After the project Indian language corpora initiative (ILCI) that was launched by Ministry of Communication and IT [4], the corpora on Indian languages are in a reasonable amount. But, still the resources available for both Hindi and Tamil language are restricted to a specific domain. The resources on Indian languages are still on a lower side based on the quantity and quality. Thus, this resource availability issue needs to be addressed so that it has lesser impact on the overall translation accuracy. This issue can be handled by using an intermediate natural language called as pivot language. The major constraint for the pivot language is – there should be vast resource availability for the pivot language.

Objective-3: To retain the semantics that has been distorted in pivot-based Hindi to Tamil machine translation system

In the pivot-based machine translation, there is loss of semantics due to the introduction of an intermediate pivot language. The pivot language has its own language specific feature which are not similar to the language specific features of Hindi and Tamil. Due to this there is loss of semantic during the translation from Hindi to Tamil using a pivot language. This requires introduction of an approach to handle semantic distortion while designing a machine translation system. One such approach is to make use of word sense disambiguation before the transfer phases. Thus, this can improve the accuracy of translation being generated.

Objective-4: To improve the accuracy of Hindi to Tamil machine translation by using syntactic and semantic features

During the introduction of word sense disambiguation in a pivot-based translation, there is introduction of noise which deviates the translated sentence away from its original target sentence in the corpus. To further improve the accuracy of translation, there was a need to explore more on the machine translation system that consider both the syntactic and semantic features of the languages being considered. The designed machine translation system for Hindi and Tamil language basically considers both the syntactic and semantic information before the transfer phase. The syntactic information is extracted from the source text with the help of Hindi part-of-speech tagger and semantic information is extracted using the word sense disambiguation phase.

Objective-5: To improve the accuracy of Hindi to Tamil machine translation without pivot language being used

The pivot language is basically introduced to handle low-resource availability issue. But, the pivot language can handle it upto certain extent only. After a particular threshold, the accuracy of the translation saturates and then starts declining. Thus, in order to handle such issue, a deep learning approach needs to be proposed to perform translation from Hindi to Tamil. This deep learning approach should be able to learn the syntactic and semantic features in the languages. The syntactic and semantic features are extracted and encoded in to a vector form with the help of word embedding phase. These vectors are further used by the proposed deep learning approach to learn about the translation features. Initially, the deep learning approach was not generating the target text in a grammatically correct manner. To handle this, an attention mechanism was introduced in the proposed deep learning approach.

1.5 APPROACH

To start with the sense-based machine translation research, the Hindi to Tamil statistical machine translation system was explored and designed. The statistical machine translation systems had made a remarkable progress in this area during the last few decades [5]. The system applies machine learning algorithms over the statistical features of the languages. The statistical features are extracted from the parallel corpus which contains the manually translated text in both the languages. The statistical machine translation system basically works on syntactic features of the words in the corpus. The proposed statistical machine translation is a modified version of the existing one such that the modification incorporates the word's semantics during the translation process. The overall system architecture is as shown in Figure 1.1. The modification is performed in such a manner that the set of words in source text provides a contextual information based on its usage and this contextual information assists the statistical machine translation system to generate the target text. The probable semantically equivalent words in Hindi are retrieved by using word sense disambiguation over the source text.

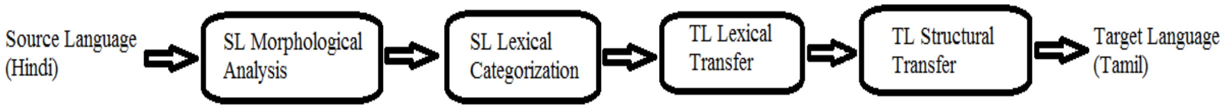

Figure 1.1: Overall system architecture for Hindi to Tamil machine translation system

The statistical machine translation system designed in the first part of research needs vast amount of resources. Unfortunately, the resource in both Hindi and Tamil language is poor. There is need for a mechanism to handle such low resource availability issue. One such mechanism is pivot based machine translation [6]. Due to the introduction of pivot language, the overall task of translation is sub divided into subtask that translate the source to pivot and then pivot to target language. The prime constraint of the pivot language is it should have vast resources due to the dependency of the overall system performance over the training data size [7]. One of the most appropriate option for pivot language is English due to the availability of vast resources for it. Thus, the Hindi to Tamil statistical machine translation system is made in two phases with the first phase translating from Hindi to English and in the second phase the generated English text is translated to Tamil language.

Introduction of pivot language in between source and target language degrades the translation accuracy due to the involvement of three different languages. Some words in Hindi will be semantically mapped to one single word in the English language and thus the semantic information that was in the source text is lost during translation through English. For example (as shown in Table-1.4), all the variants of Hindi word that is referred as “aunty” in English, but, each of these words have individual equivalent Tamil words. To handle this kind of words, a word sense disambiguation module was included while designing the hybrid machine translation system.

Table 1.4: Example of many-to-one mapping between Hindi and English

S. No.	**Hindi Word**	**English Word**	**Tamil Word**
1	चाची (chaachii)	Aunt	சித்தி (chiththi)
2	ताई (taaii)		பெரியம்மா (periyamma)
3	बुआ (buaa)		அத்தை (Attai)
4	मौसी (mausii)		சித்தி (chiththi)
5	मामी (maamii)		அத்தை (Attai)

The hybrid machine translation system that makes use of word sense disambiguation on a pivot-based system saturates after a certain threshold and after a certain limit beyond saturation, the accuracy starts to decline. Thus, to improve the accuracy of Hindi to Tamil machine translation, the neural machine translation was designed which uses the syntactic and semantic features of the language. The neural machine translation is fed the syntactic and semantic features of the languages in the form of word vectors. The word vectors are generated by the feature learning technique that maps the words to a vector of values. The word vectors are generated using the word embedding module of the neural machine translation system. These vectors are fed to neural machine translation system so that it can learn translation between the texts fed to it. The translation generated by the neural machine translation system lacks in the target language grammatical correctness. Thus, an attention mechanism was introduced in the neural machine translation system to take care of grammatical errors.

1.6 CONTRIBUTIONS

Based on the challenges mentioned in section-1.3 and to be in accordance with the main objective of this research, a novel approach for sense-based Hindi to Tamil machine translation was designed and developed. The novel approach for sense-based Hindi to Tamil machine translation has the following significant contributions from this research work,

1. The first and primary contribution of this research work is in designing a machine translation system that makes use of both the syntactic and semantic features of the languages. These features are extracted from the parallel bilingual corpus that is being used in the domain of research, machine translation. The syntactic and semantic features are extracted from the parallel corpus with the help of part-of-speech tagger and word sense disambiguation modules respectively. During translation, these features are used to predict the probable target language text. To translate an input word in Hindi to its corresponding Tamil word, the transfer phase was developed using the three factors - word's positional information, word's part-of-speech and the word's semantics. Due to these factors, the translation quality gets enhanced and thus benefits the overall system.

2. Word alignment models align the words in source text with its corresponding words in target text using the bilingual corpus. The second main contribution is in modifying the word alignment model for Hindi and Tamil language pair since the existing word alignment models are not sufficient for both these languages. The Hindi language has multiple auxiliary verbs along with the main verb of the text, whereas, it is not the case with Tamil language. For example, the phrase "tej kar dete hain" mentioned in Table-1.5 has the main verb as "tej" and the rest of words in the phrase are auxiliary verbs for this main verb. The statistical machine translation system's performance is based on the accuracy of these word alignment models, due to which the need has increased to incorporate the syntactic nature of languages in the

alignment models. The existing word alignment model is modified such that the part-of-speech of the words also has a role in the alignment process.

Table 1.5: Example phrases showing usage of auxiliary verbs

S. No.	Hindi Phrase	English Equivalents	Tamil Phrase
1	तेज कर देते हैं । (tej kar dete hain)	Accelerate	அதிகமாகிறது (Atikamākiṟatu)
2	बढ़ता है । (badhata hai)	Increase	அதிகரிக்கிறது (Atikarikkiṟatu)
3	साफ करें । (saaph karen)	Clean	சுத்தம் செய்யுங்கள் (Cuttam ceyyuṅkaḷ)

3. The second main contribution is to handle the low-resource availability issue in both Hindi and Tamil language. One such approach to handle the low resource availability issue is by using a natural language as intermediate pivot language in between Hindi and Tamil. Thus, English being a vast resource language, it has been chosen as the pivot language.

4. The next major contribution in this research is in designing a hybrid system which handles the low resource availability issue without compensating on the overall translation quality. In this hybrid approach, a pivot language based statistical machine translation was introduced and in order to handle the semantic distortion that occurred in pivot-based system, the word sense disambiguation module was introduced to identify the word’s sense. This hybrid approach helps to improve the translation quality which got degraded when pivot language is introduced.

5. The last and important contribution in this research work is to design a deep learning-based approach for the machine translation. While designing this deep learning-based system, both the syntactic and semantic features are embedded so that the overall translation quality is not

compensated. This deep learning-based approach gets trained with the training corpus and learns about the features of the training data. Since the syntactic and semantic features are embedded in the vector, the deep learning approach learns based on these features as well. Thus, a sensible translation is generated by the deep learning approach.

Further, the deep learning approach should generate the output text in accordance with the grammar in Tamil language. This is also considered while designing this deep learning-based system and an attention mechanism has been applied to achieve this.

1.7 THESIS ORGANIZATION

The thesis has been organized into seven chapters. The brief outline of each chapter is given below:

The chapter 1 is introduction of the subject. It discusses the present scenario in machine translation system. It also discusses about the motivation, objective, issues and solution approaches used in the present research.

Chapter 2 focuses on existing related work on machine translation systems. This chapter discusses about various approaches that have been used in a machine translation and its related merits and demerits. It also discusses about the existing Indian language processing systems that can aid in translation process and the various issues being handled in such systems.

In chapter 3, the focus is on the word sense-based Hindi to Tamil machine translation system. It describes about the phases used in the word sense-based statistical approach and discusses about the results generated using this approach. This chapter has the description about the short-comings in this approach and also proposes the approach to handle such short-comings in brief.

Chapter 4 deals with pivot-based Hindi to Tamil machine translation system. It details about the need for a pivot language and how to choose a pivot language based on various factors. Also describes about the phases used in this approach. The chapter gives the analysis of results generated along with the demerits of this system and methods to handle such demerits.

In chapter 5, a hybrid machine translation system has been introduced which makes use of both the word sense as well as the pivot language. This chapter describes about the phases and working of the system. The result analysis of this approach is described in this chapter along with its short-comings with respect to the two languages under consideration, Hindi & Tamil. It briefly describes about the ways to handle such short-comings.

Chapter 6 introduces the deep learning approach for Hindi to Tamil machine translation. It details about the need for such a system and various stages that will be used in this approach. Each stage used in this approach has been described in detail. Results of this approach have been analyzed and compared with the preceding chapter results.

Chapter 7 summarizes the contributions of this research work and indicates the scope for the possible future extensions in this area of research.

1.8 CONCLUDING REMARKS

This chapter introduces to the area of machine translation and discusses in detail about the various issues involved in machine translation developed for Indian languages. It also describes about the various solution approaches for handling these issues. This chapter also describes about how these approaches help in improving the overall translation accuracy. These solution approaches have been designed using four different system and it has been briefed in this chapter. The rest of the thesis describes in detail about these four systems along with its result analysis. The next chapter is about various related work that has been carried out by various researchers across the world in this domain of research.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 INTRODUCTION

This chapter describes about the various related works carried out all over the world in machine translation. It also details about the research works performed specifically on Indian languages. Also, there is discussion about various research gaps identified in the existing machine translation system. It briefly details about the ways to handle such research gaps so that the performance of the machine translation system can be improved. The chapter is further divided into four sections and the details about each of the sections is as follows:

Next section provides an overview of various machine translation systems developed over years for Indian languages. It summarizes about the various approaches that was used in each of the highlighted machine translation systems, besides providing detailed description of a few important machine translation approaches. Special attention has been given to existing machine translation system that uses one of the languages as Indian language.

Section-2.3 highlights about recent works that have been carried out on machine translation. In general, the section describes machine translation in a broader way without any restriction on the natural language. It also describes the approaches being used in these systems, along with the merits and demerits of these approaches.

Based on the survey of the related works mentioned in section-2.2 and section-2.3, the identified research gaps are described in section-2.4. This section also describes in detail about the research gaps on machine translation system that was developed by keeping Indian language features into consideration. Also describes about the impact of these research gaps on an Indian language machine translation system and proposes the approaches that can be used to handle it.

Last section of this chapter concludes about the highlighted works that was proposed by various researchers and also briefly discusses about the method that can be used to handle the research gaps without compromising on the accuracy of translation.

2.2 OVERVIEW OF INDIAN MACHINE TRANSLATION SYSTEM

Machine translation systems for most Indian languages are characterized by use of any existing machine translation approaches. These approaches are broadly classified as – Rule based machine translation (RBMT), Empirical based machine translation (EBMT), Hybrid machine translation and Neural machine translation (NMT) as shown in figure 2.1. Rule-based machine translation system employs machine learning algorithm to learn the various features of the training set and these features assists the translation system later on. This kind of translation approach is further categorized into direct machine translation, transfer based machine translation and interlingua machine translation.

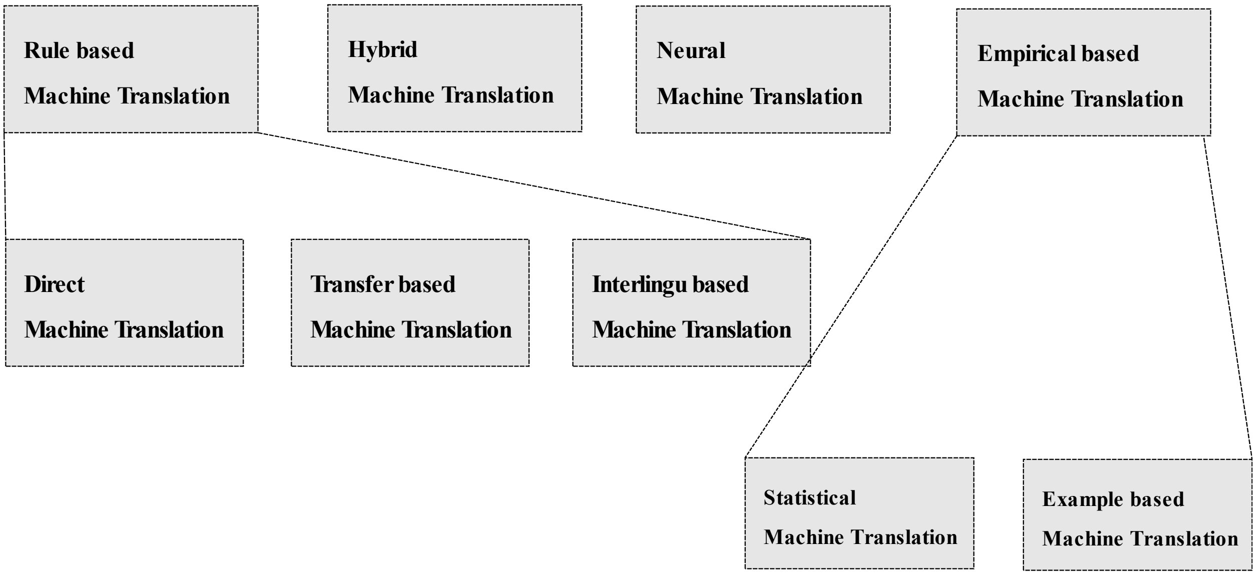

Figure 2.1: Various approaches on machine translation

Each of these rule-based machine translation systems has their own merits and demerits. The ease of implementation is an advantage of direct machine translation and it also out-performs others when the languages being considered are grammatically related [8]. The overall architecture of a direct machine translation is as shown in Figure-2.2. In this approach, the source word is translated to its respective target word using the bilingual dictionary. The bilingual dictionary consists of rules of mapping the source word with target word. Later, the translated text is rearranged to make it grammatically correct in the target language. Its demerits include,

i. Relationship between words in text are not considered in a RBMT system
ii. Loss of semantics during translation
iii. Quite expensive for multilingual scenarios [9] and it is not adaptable to different language pairs

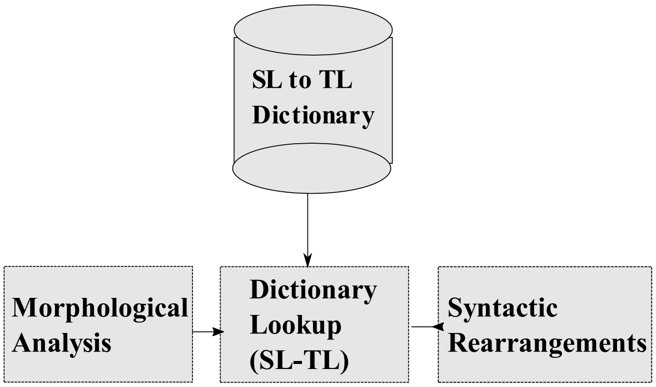

Figure 2.2: General architecture of direct machine translation approach [8]

The general architecture of transfer-based approach is shown in Figure-2.3. The main advantage of interlingua and transfer based machine translation is its modularization into sub-modules that handles the translation in parts. The first module translates the source language to intermediate language and then second module translates the intermediate language to target language. Due to modularization, this machine translation approach is cost effective when compared with direct machine translation approach. Both these approaches resolve the ambiguities between one language to other. The major disadvantage of both these approaches is semantic loss during the translation process. The intermediate representation in both these approaches is also a challenge since the representation should carry the semantics of source so that the target text is semantically aligned with respect to source.

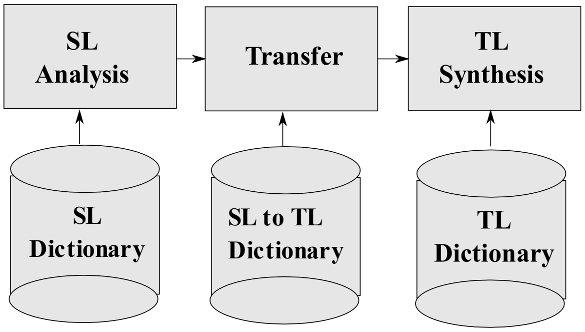

Figure 2.3: General architecture of transfer-based machine translation approach

Empirical-based machine translation system uses machine learning algorithm to learn the statistical features of the languages and these features help in the translation process. The empirical based machine translation is sub-divided as statistical machine translation and example-based machine translation. The statistical approach makes use of the statistical features which are extracted from the target language word order and source-target word pairs. This approach makes use of a statistical model developed using these features [10] and the general architecture of statistical machine translation system is as shown in Figure-2.4. These statistical features are extracted from the bilingual parallel corpus. One of the major advantage of statistical approach is its adaptability. So, it can be used for any language pairs. The following are the list of disadvantages of statistical machine translation approach,

i. It has a great dependency with parallel bilingual corpora.

ii. Creation of parallel corpus from the available limited resources is very hard

iii. The performance of the statistical system has a dependency with the quality of corpus being used.

iv. It is not a suitable approach for languages with different word orders.

v. Has dependency on the word alignment models which will be used to extract the statistical features

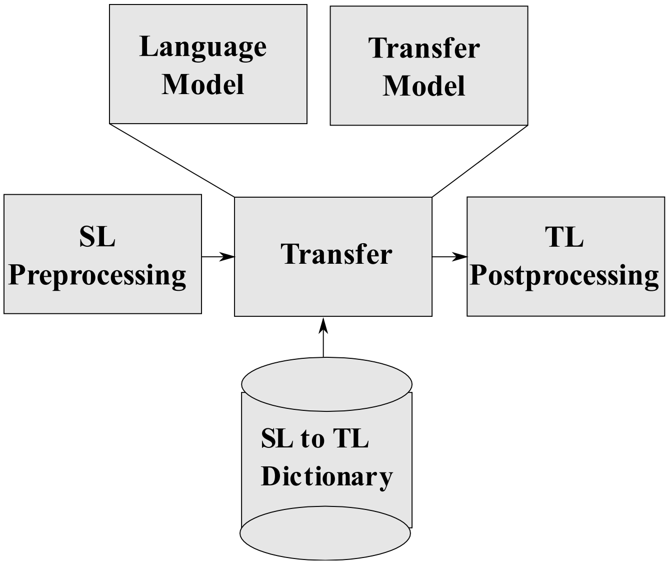

Figure 2.4: General architecture of statistical machine translation approach

Example based machine translation approach analyzes the input sentence and maps it with the example source sentence in the corpus provided to it. Once there is a matching sentence found in the corpus, it performs the translation according to the target example sentence given in the corpus. EBMT is as shown in Figure-2.5. The retrieval module should be more precise and in general, both the syntactic and semantic similarity measures are used to perform the matching.

The main demerit of example-based machine translation approach is that its inefficiency to handle same type of source sentence that are being translated to structurally different target sentence. This is also called as translation divergence [11]. From example, consider structurally similar sentences – “he is happy”, “he is busy” and “he is scared”, which has translation divergence with its equivalent Hindi translations. The sentences “he is happy” and “he is busy” has the Hindi translation as “vah (he) khush (happy) hai (is)” and “vah (he) vyast (busy) hai (hai)”. But, in case the sentence “he is scared”, the translated Hindi text will be “vah (he) ghabaraaya (scare) hua (-ed) hai (is)”. Due to this variation, if the example target text chosen for translation is not proper, then it will degrade the performance of translation. In general, the example-based machine translation system is designed by giving priority to translation divergence.

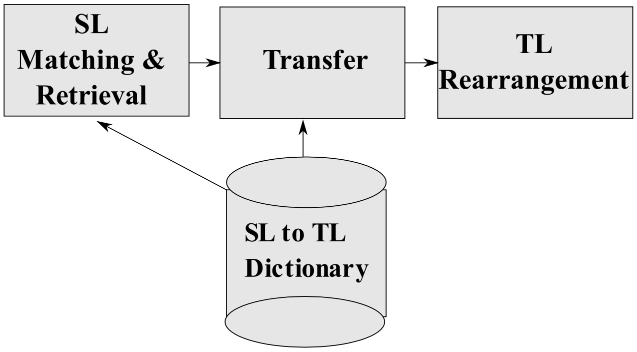

Figure 2.5: General architecture of example-based machine translation approach

Hybrid approach can have any of the combinations of these approaches based on necessity of the system. The features of various approaches are exploited to choose the appropriate combination for developing a hybrid machine translation system. The choice of approaches has a great dependency with the language specific features. There are many existing machine translation systems that makes use of hybrid approach such as AnglaHindi [12], English to Devanagari translation [13], lattice based lexical transfer in Bengali-Hindi language pair [14]. In general, these existing hybrid machine translation systems uses multiple machine translation engines to generate probable translations and all these generated translations are combined to generate the final translation.

Neural machine translation system makes use of deep learning networks to extract the features of languages and map it according to the parallel sentences fed to the system. In general, the neural machine translation system has two sequential stages known as encoder and decoder as shown in Figure-2.6. Encoder basically embeds the variable length input sentence to a fixed length intermediate vector. The decoder processes each word at a time from the intermediate representation and generates its target sentence [15].

As discussed in this section, each of these existing approaches has their own merits and demerits. The choice of the approach has a strong dependency with the languages being chosen for the machine translation. This is due to the language specific features that affect the performance metric of various approaches. Still there is need for modification on these existing approaches, so that the modification can refine the demerits of these approaches.

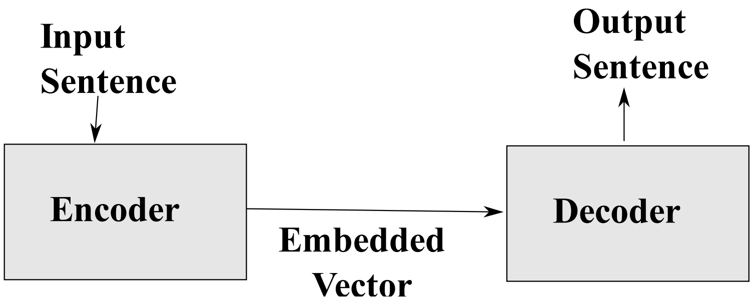

Figure 2.6: Neural machine translation approach

There are various existing Indian language machine translation systems that makes use of these approaches that have been discussed. One of the Indian language machine translation systems that uses direct word to word translation is – Hindi to Punjabi machine translation system developed by Vishal Goyal et. al [16][17] in the year 2009, 2011. To improve the grammar of the translated text, this direct machine translation uses a post processing phase which was developed using regular expression and pattern matching algorithm. This system's accuracy has been reported to be around 95.4% and Bilingual Evaluation Understudy (BLEU) score as 0.7804. This is one kind of direct machine translation system and there are other systems too such as – Anusaaraka (1995), Web based Hindi to Punjabi machine translation (2010). The accuracy of these systems is in the range of 80% to 95%. Shachi Dave et. al [18] proposed English to Hindi machine translation using interlingua approach. The intermediate language being used in this approach was universal networking language (UNL) and this intermediate language acts as a network of words with its semantic relations.

The existing Indian language machine translation system that uses empirical based machine translation approach – English to Indian language machine translation (E-ILMT) that was developed by nine Indian institutions in the year 2006 [19]. It basically uses statistical machine translation approach and it also uses morphologically processed text during the training of the statistical machine translation system. Before preprocessing, the text is subjected to syntactic reordering. So that, it reduces the number of rearrangements that will be required during translation. Anoop Kunchukuttan et. al. [20] proposed an enhanced phrase based statistical machine translation by applying source level reordering which is followed by transliteration of words that are not translated words. Prof. R M K Sinha et. al. [21] has developed ANGLABHARTI-II, a generalized example-based approach to perform machine translation. This approach initially searches for a match in the example base that was developed for languages under consideration. If there isn’t a match found for it, then the system invokes the rule base to perform translation. A transfer-based English to Hindi machine translation (MANTRA) was developed by Hemant Darbari et. al. [22]. This transfer-based system makes use of tree adjoining grammar (TAG) and also lexicalized adjoining grammar to represent the grammars in English and Hindi. Later the tree like structures are used for translating the text. A hybrid approach to perform translation from Bengali to Hindi was proposed by Chatterji S et. al. [23]. This hybrid approach is proposed by integrating a statistical machine translation with a transfer-based approach. The BLEU score of this system is found to be better when compared with the BLEU score of systems that uses these approaches individually.

2.3 RELATED WORKS

For English to Hindi statistical machine translation, the key challenge is that the Hindi language is richer in morphology than the English language. There are two strategies that facilitate reasonable performance in this language pair. Firstly, reordering of English source sentences in accordance with Hindi language and the second strategy is by making use of suffixes of Hindi words. Either of these strategies or both the strategies can be used during translation. The difference in word order of Indian language and English, makes these two strategies as challenging one. For example, the English sentence – “he went to the office” and its corresponding Hindi translated text is “vah (he) kaaryaalay (office) gaya (went to)”. In this example, it is evident that the position of words in English is not retained in the translated Hindi text. Since Indian languages are morphologically rich and has limited availability of parallel corpora, the above two strategies are more challenging to derive the desired results with reasonable performance [24]. The authors have proposed one such statistical machine translation system which makes use of both these strategies. It is a phrase based statistical machine translation system. The morphological information of both the Hindi and English languages are extracted and fed to the word alignment models. The output from the word alignment models is fed to a phrase extraction phrase that constructs a phrase table. Using this phrase table, the translation between these languages are performed. The results of this system show that there is increase in the BLEU score of the system by the introduction of both the syntactic information and morphological information.

Ali Hasan Imam et al has experimented on English-Bangla statistical machine translation with different corpus size and has provided with various observation on the change in corpus size [25]. The author has observed that the translation quality is improved when there is increase in corpus size. But, there is saturation point after which increase in corpus size has least impact on the translation accuracy. The author has suggested that this can be handled by improving the quality of corpus after the saturation point. Authors have built their own corpus from different sources on these two languages and has assessed the machine translation framework in the light of BLEU score.

Felipe Sanchez-Martinez et. al [26] proposed an efficient Hidden Markov Model (HMM) [27] for performing the part-of-speech tagging. In this proposed system, the author has developed an unsupervised part-of-speech tagger that makes use of target language information and it has been proven that the results are better as compared with Baum-Welch algorithm [28]. The main demerit of this approach is the increase in the required number of translations in an exponential manner. Thus, increasing the overall time of execution due to the increase in time for translation. To improvise this algorithm, the author has applied pruning method to identify the unlikely disambiguation. Thus, reducing the number of translation required and also the overall time complexity of the algorithm.

In 2014, Aneerav Sukhoo et. al. has explored about statistical machine translation amongst English and Mauritian Creole language pairs. It has been developed for tourism and business domains [29]. In the development of this statistical machine translation system, author has used MOSES tool [30] and found that performance was not as expected which is due to very small parallel corpus. Additionally, author has also utilized the bilingual dictionary on English-Mauritian creole language pairs to enhance its performance. The author has observed that the BLEU score to be roughly around 6.0 in the scale of 0-10. It has been observed that the BLEU score is directly proportional to the corpus size and thus the increase in corpus size also increases the BLEU score of the system.

A novel neural network based bilingual language model was proposed by Rui Wang et. al. for Chinese-English statistical machine translation system [31]. A continuous space language model utilizes a monolingual corpus to generate the language model. The author has introduced a method to modify a continuous space language model to a bilingual continuous space language model. Using one million parallel training corpora on Chinese-English language pairs, machine translation system was developed which utilizes various language models over Chinese-English. To reduce the computational and space complexity in the phrase processing module, the author has proposed an approach to use only top-rated phrases which are identified using sentence ranking method. This proposed system on statistical machine translation performed very well over the various existing language models by either converting or growing methods.

In 2015, DeyiXiong et al has proposed a document level topic-based coherence model for statistical machine translation [32]. From the source text and target text of the parallel corpus, coherence chain was extracted for both the languages under consideration. Using maximum entropy classifier, the source coherence chain is further matched on to the target coherence chain. The author has developed two variants of the topic-based coherence models – the first one at the word level coherence model and later one on the phrase level. The performance of phrase level system is found to be better as compared with the word level system. This overall performance of the developed system is also better when compared with the baseline system.

The authors, Cyrill Goutte et. al, has defined machine translation in terms of two sub-tasks namely, lexical selection and lexical reordering [2][33]. The lexical selection to basically to choose the appropriate target language word for the given source word based on the identified features. In a lexical ordering, the chosen target words are re-ordered to generate a meaningful and grammatically correct target sentence. The word alignments models that was used in GIZA++ [34], considers the local association between the words in source and target language. But in certain conditions, the association is not locally bounded. Thus, to further improve, the word alignment algorithm can be improved by using phrase alignments between the source and target texts.

Di He et. al [35] has proposed an approach to improve the machine translation by using dual learning mechanism. This approach has been developed to improve the overall accuracy of the system with the help of the monolingual corpus. In this approach, there are two agents being used to handle task of translation from source to target and vice versa. These agents are further coordinated using reinforcement learning to improve the learning from the training corpus. The process of reinforcement learning is being continued further until the accuracy of both the agents converge.

In Sutskever et. al, sequence to sequence learning with neural networks was developed, which consists of an encoder and a decoder [36]. Encoder converts the input sentence into a fixed-length vector. These vectors are further converted into target sentences by the decoder module. Both these encoder and decoder are jointly trained to enhance the probability of correct translation. The issue that occurs during training a neural network is overfitting. A simple way to prevent overfitting in neural network was introduced by Srivastava et. al [37]. In this paper, the author introduced an approach called as dropout regularization. During the training of neural network, the weights of neuron are randomly made zero based on the dropout percentage which prevents the neurons from adapting too much according to the training data. Author has also found that it improves the problem of overfitting in significant manner as compared with the other regularization methods.

Shahnawaz and Mishra [38] have presented English to Urdu machine translation system which makes use of case-based reasoning (CBR) to identify the translation rules and these translation rules are further used in the artificial neural network. The case-based reasoning technique is used as learning process to extract the translation rules for the Urdu sentence from the input English sentence. These translation rules are further used in the artificial neural network and the author has found that the system’s performance as 0.728 BLEU score which is good for those sentences for which the cases are available in the case base.

Bakhouche et al. [39] have proposed an ant colony optimization algorithm for word sense disambiguation of Arabic words using the lexical information of corresponding English words in Princeton WordNet. The Arabic WordNet has been mapped with the Princeton WordNet and this mapping is used to identify the lexical information of the word under consideration. The semantic information about the words are extracted both locally and globally. The semantic information extracted locally with the help of Lesk algorithm [40], [41] and are propagated globally using the ant colony optimization algorithm. The performance of this system is found to be approximately 80%.

Soltani and Faili [42] have developed a system to identify the probable target word for the given source word. A bilingual dictionary is being used to retrieve the various possible target words and then the most appropriate probable target word was chosen using the semantics. The proposed system also generates the semantic dependency graph of different sense of word and ranks each node/edge in the graph. This semantic dependency graph is used to identify the target word in English to Persian machine translation system. This system has shown considerable improvement as compared with other word sense disambiguation methods.

Neural machine translation's accuracy depends more on the encoder which captures the semantic information that is being conveyed in the source text. These encoders can be unidirectional or bidirectional recurrent neural network (RNN). In the paper authored by Biao Zhang et. al., the author has proposed a context aware recurrent neural network for the encoders in a neural machine translation. This context aware encoders works in two level hierarchy, where history information is extracted in the bottom level and are aligned with the future context in the upper level [43].

There are various deep neural networks designed by researchers such as deep belief networks, recurrent neural network, convolutional neural networks and deep stack networks. Jiajun Zhang et. al., has mentioned that there are various types of deep neural networks being used in language modeling, word alignment, translation rule selection and reordering which are main phases of a machine translation system. Apart from having different deep neural network for each of these tasks, it is also better to have a pure neural machine translation system since deep neural networks facilitates the computation of semantic distance between texts. The issues that will be existing in machine translation system with deep neural network are – computational complexity, error analysis and reasoning [44].

Since sense-based machine translation is under progress these days, there is multi-sense based neural machine translation being introduced by Zhen Yang et. al. [45]. In their proposed system, they generate the word embedding based on different senses of the word, called as sense-specific embedding. The multi-sense embedding module was developed using a recurrent neural network (RNN). Zhen Yang et. al. also proposes that this sense-based embedding is task independent and it can be applied on any natural language processing task.

2.4 RESEARCH GAPS

Based on the literature survey, the following issues were identified which motivated to propose the solution in this research work

i. Machine translation system for Indian language pairs has scope for improvement on the accuracy by considerable amount

ii. The existing systems do not consider the contextual information of the words during translation

iii. Necessity for modification on existing word alignment algorithms to handle features of Indian languages

iv. In general, the resources for Indian languages are limited.

v. As compared to English, Hindi and Tamil are free word order languages.

vi. Hindi and Tamil are morphologically rich languages

vii. Semantics are lost during translation from Hindi to Tamil using English as a pivot.

Issue 1: Machine translation system for Indian language pairs has scope for improvement on the accuracy by considerable amount.

Indian languages are free word order languages and morphological structure of words in these languages are different as compared to English. Few examples of free word order in Hindi and Tamil is shown in Table 2.1. The existing machine translation approach for English language will not be suffice for the Indian languages. Morphological structure of words in Hindi and Tamil has information such as TAM (tense, aspect and modality). For example, as shown in Table 2.1, the word ‘padhate’ in Hindi has both the gender information as well as plural information attached with the root word ‘padh’. The word ‘paṭikkiṟārkaḷ’ in Tamil also has gender and plural information attached with the root word ‘paṭi’. In case of English the gender information is not attached along with main verb. For example, the main verb “read” and its variants “reads”, "reading", "read" does not contain any gender information. This TAM information plays a vital role during translation process. In certain cases, this TAM information of main verb is available in the auxiliary verb mentioned in the text. For example, the phrases "padh raha hai" and "padh rahi hai" has difference in gender. Similarly, the other TAM information is also provided in the auxiliary verb mentioned. The TAM information can be detected using morphological analysis and part-of-speech tagger methods. But the accuracy of these methods on Hindi and Tamil language is still poor compared to other languages such as English. Thus, the accuracy of translation system also degrades with respect to decrease in accuracy of morphological analysis and part-of-speech tagger in Indian languages.

Table 2.1: Few examples to highlight free word order feature in Hindi and Tamil

S. No.	Hindi & Tamil Sentence	English Equivalents	Reordered Hindi & Tamil Sentence (Single variant)
1a	बच्चे किताब पढ़ते हैं । (bachche kitaab padhate hain)	Children are reading books.	किताब पढ़ते हैं बच्चे। (kitaab padhate hain bachche)
1b	குழந்தைகள் புத்தகத்தைப் படிக்கிறார்கள். (Kul̲antaikaḷ puttakattaip paṭikkir̲ārkaḷ.)		குழந்தைகள் படிக்கிறார்கள் புத்தகத்தைப். (Kul̲antaikaḷ paṭikkir̲ārkaḷ puttakattaip.)
2a	हमे सेब के पेड़ से सेब मिलता है। (hume seb ke ped se seb milata hai.)	We get apple from apple tree	हमे सेब मिलता है सेब के पेड़ से । (hume seb milata hai. seb ke ped se)
2b	ஆப்பிள் மரத்திலிருந்து ஆப்பிள் கிடைக்கும். (Āppiḷ marattiliruntu āppiḷ kiṭaikkum.)		ஆப்பிள் கிடைக்கும் ஆப்பிள் மரத்திலிருந்து. (āppiḷ kiṭaikkum.Āppiḷ marattiliruntu)

3a	पानी पीना स्वास्थ्य के लिए अच्छा है। (paanee peena svaasthy ke lie achchha hai.)	Drinking water is good for health.	पानी पीना अच्छा है स्वास्थ्य के लिए । (paanee peena achchha hai svaasthy ke lie.)
3b	குடிநீர் ஆரோக்கியத்திற்கு நல்லது. (Kuṭinīr.ārōkkiyattiṟku nallatu)		நல்லது ஆரோக்கியத்திற்கு குடிநீர். (Nallatu ārōkkiyattiṟku kuṭinīr.)

Issue 2: The existing systems do not consider the contextual information of the words during translation.

In Hindi and Tamil languages, sense of the word changes according to the context where it is being used. For example, the word 'Aam' in Hindi has different sense in both these phrases 'Aam aadhmi' (common man) and 'Aam ka ped' (mango tree). In 'Aam aadhmi' it is talking about the common man whereas in case of the phrase 'Aam ke ped' it refers to 'Mango tree'. The word 'Pooja' in Hindi has different senses according to its usage. 'Pooja' can refer to name of a girl and it also refers to worshiping god. Similarly, in Tamil language the word 'padi' has different sense according to the context it is being used. In phrases 'puttakam padi' and 'padi mel yeriva'. In 'puttakam padi', the 'padi' refers to the read whereas in the other phrase it refers to the steps. In case the contextual information is not used during translation, the meaning of the translated sentence may get distorted and the actual information which was conveyed in the source text won't be available in the target text. Thus, the contextual information plays a vital role too.

Issue 3: Need for modification on existing word alignment algorithms to handle features of Indian languages.

The word alignment models are used to align the words/phrases in source text with its corresponding words/phrases in the target text. The accuracy of the machine translation has a dependency with the accuracy of word alignment model being used [46]. Brown et. al [47] has proposed statistical models on word alignments with different variants based on its number of features being used. These existing word alignment models are not sufficient for languages like

Hindi and Tamil. This is due to the language specific features such as morphological richness, free word order and paired words. There is need for some modification that can handle these features, so that the overall translation accuracy is not compromised.

Issue 4: In general, the resources for Indian languages are limited.

The resource availability for Indian languages are still poor [48] compared to English and other European languages. Technology development for Indian languages (TDIL) has taken initiative to develop the various resources for Indian languages. But still the resource availability is poor due to the existence of resources on few prominent Indian languages such as Hindi. Because of the poor resource availability, there will be higher impact on the accuracy of the overall system. There is need for natural language processing systems that can excel on resource constraint as mentioned in [49] by Prof. Pushpak Bhattacharya. In this research, the author has discussed about the approach to handle word sense disambiguation when there is scarcity of resources.

Issue 5: As compared with English, Hindi and Tamil are free word order languages

English has a fixed grammatical sequence as compared with Indian languages. Since Hindi and Tamil are free word order languages, the sentences in these languages are grammatically valid irrespective of the order of words. The existing state-of-art methods that was being used in English language gave poor accuracy when applied on Indian languages due to its free word order nature. Thus, there is increase in need for a better machine translation approach for Indian languages such as Hindi and Tamil.

Issue 6: Hindi and Tamil are morphologically rich languages

Since both languages are morphologically rich, there is more information being conveyed in each of the words. This information is tense, aspect and modality (TAM). Apart from these, there are multiple auxiliary verbs in Hindi which are being mapped to a single Tamil word as shown in Table-2.2. Thus, increasing the complexity of translation between these two languages. For example, the phrase “Kar raha hai” in Hindi is mapped to a single word “Ceykiṟār” in Tamil.

Table 2.2: Few examples showing need for TAM information stored in Hindi and Tamil texts

S. No.	Hindi texts	English Equivalents	Tamil texts	TAM Information
1	कर रहा है (Kar raha hai)	is doing/has been doing	செய்கிறார் (Ceykiṟār)	Present & Masculine
2	कर रही है (Kar rahi hai)		செய்கிறாள் (Ceykiṟāl)	Present & Feminine
3	करेगा (Karega)	will do	செய்வேன் (Ceyvēṉ)	Future & Masculine
4	करेगी (Karegi)		செய்வாள் (Ceyval)	Future & Feminine
5	कर लिया (Kar Liya)	have done	செய்தேன் (Ceytēṉ)	Past Tense

Issue 7: Semantics are lost during translation from Hindi to Tamil using English as a pivot

The semantic information that is being conveyed in the source text (Hindi) will be lost in case if English is used as an intermediate language for the translation from Hindi to Tamil. Certain words in Hindi has a one-to-one mapping with the respective Tamil words whereas the mapping with English has a many to one or one to many relationships as shown in Table 2.3. This leads to loss of semantics during the Hindi to Tamil machine translation using English as the pivot language. Thus, the need for a modified approach that considers the word's sense during translation has increased.

Table 2.3: Few parallel examples of Hindi-English-Tamil texts

S. No.	Hindi Text	English Text	Tamil Text
1	खाना (Khana)	Food, Eat	சாப்பிட (Cāppiṭa)
2	पूजा (Pooja), उपासना (Upasana), आराधना (Aradhana)	Worship	வழிபாடு (Vaḻipāṭu), பூஜா (Pooja)

2.5 CONCLUDING REMARKS

Based on the analysis on various related work, there are certain research gaps that need to be addressed in order to have a more accurate translation. The research gaps identified are with respect to the Indian languages being considered. To address these issues, there is need for language specific models to handle it. With respect to a statistical machine translation system, there is need for additional phase which has to be incorporated in the source language analysis phase just before the transfer phase. Also, to perform the statistical machine translation between Hindi and Tamil, there is need for improvement in the word alignment models. These word alignment models should be such that it can handle the language specific features. To handle the low resource issue, there is need for an intermediate pivot language between the source and target language. Based on the literature survey, it is identified that the pivot language that can be used in between should have vast resources. The solution approach to handle the identified issues are described in detail in the next chapters. The performance of the identified solution approach is analysed and compared in each of the chapters.

CHAPTER 3

WORD SENSE BASED STATISTICAL MODEL OF MACHINE TRANSLATION

3.1 INTRODUCTION

This chapter addresses the word sense-based approach for Hindi to Tamil statistical machine translation. Various phases of processing involved in word sense-based approach are shown in Figure-3.1. The first phase is pre-processing that is required before actual translation. Pre-processing is performed on the words mentioned in the input text to separate the affixes attached with the root word. The extracted affixes have tense, aspect and modality (TAM) information. This information also helps in transfer phase for generating a meaningful translation. After the pre-processing is performed, the following are three main phases of the word sense-based approach:

i. Source language analysis
   a. Part-of-speech tagger
   b. Word sense disambiguation
ii. Alignment phase
iii. Transfer phase

The main objective of this system is to develop a statistical machine translation system which considers the syntactic and semantic features of the language being used. The syntactic and semantic features are extracted from the source language text and are used in the transfer phase to generate the sentence in the target language. This information is extracted by the source language analysis phase. This phase is further sub-divided into – part-of-speech tagging and word sense disambiguation. Once this information is extracted, it is used in the transfer phase which makes use of Bayesian approach. This sense-based machine translation system makes use of Bayesian approach, which is based on the statistical analysis of existing bilingual parallel corpora. The statistical analysis is also performed based on the local syntactic information available in the input sentence. Using the analysed data, the Bayesian approach is applied to predict the probable target word for the given input word. This proposed word sense-based statistical machine translation system may be mathematically expressed as,

$$P\left(\frac{W_t}{W_s,Tag_s}\right) = P\left(\frac{W_s,Tag_s}{W_t}\right) * P(W_t) \tag{3.1}$$

Where,

$W_s - Source\ word$

$Tag_s - Part - of - speech\ of\ source\ word$

$W_t - Target\ Word$

The probable target word W_t has dependency on the source word W_s and its part-of-speech Tag_s. Thus, to find the probable target word W_t with the prior knowledge about the input source word and its predicted part-of-speech, a Bayesian approach is used. The Bayesian approach employs two models to predict the probability of target word. The two models required are translation model $P\left(\frac{W_s,Tag_s}{W_t}\right)$ and language model $P(W_t)$. The translation model is used to find the probable source word for a given target word. In the language model, it predicts the probable target language word at nth position based on the predicted (n-1) target words [2].

Since the language model has dependency on target language, there is need to restructure the source text based on the grammar of the target language. This rearrangement of words is performed in the alignment phase and it must be performed before the transfer phase due to the need for target language grammar in the language model.

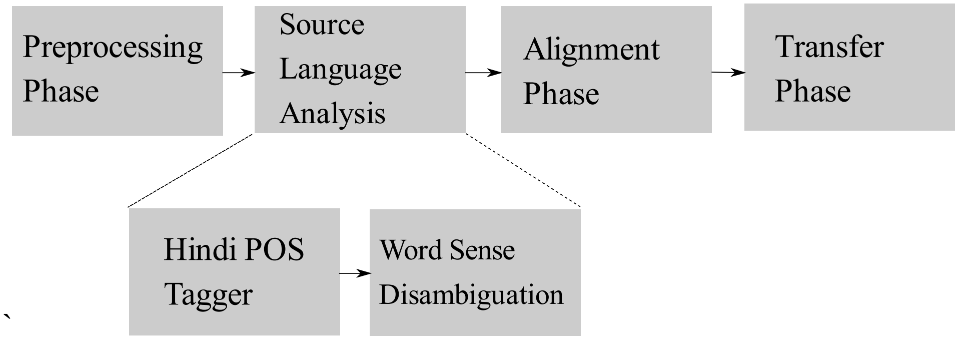

Figure 3.1: Architecture of proposed word sense-based statistical approach

3.2 FEATURE EXTRACTION

To develop a statistical machine translation system, there is need for statistical information about the mapping between the languages. As said earlier, in the transfer phase of word sense-based approach, there is need for a translation model and a language model. The statistical information about the mapping between source word and target word is required for the translation model. Whereas, in the language model, the statistical information about mapping between current target word and its previous target word is required. Both the statistical information is retrieved using the syntactic feature of the languages as well. In the proposed word sense-based approach, the syntactic information and semantic information are passed on to the alignment phase and transfer phase. Thus, the mapping that is required in both the language model and transfer model is performed using the word and its syntactic feature.

3.2.1 CORPORA USED

To extract the statistical information in the preliminary phase, there is need for parallel textual corpus on Hindi and Tamil languages. This parallel corpora was provided by Technology Development for Indian Languages (TDIL) programme, Department of Information Technology (DIT), Ministry of Communication & IT, Government of India. The corpus that is used to develop this translation system is on the health domain. There were around 25000 parallel bilingual texts in the corpus provided.

3.2.2 STATISTICAL DATA GENERATION

The statistical data required for the translation model and language model will be generated by mapping the data in parallel corpus provided by TDIL. There is need for two different statistical information. Firstly, the statistical information about the mapping between source word with its part-of-speech and its equivalent target word is generated. This mapping is performed by using a modified IBM model in which the part-of-speech of the source word is considered during the mapping with its semantically equivalent target word. Secondly, the statistical information for the mapping between the words in target language is generated. This also is done by using the modified IBM model to map the current word in target language with its preceding word. Both these statistical data are required in the alignment phase and transfer phase.

The word alignment algorithm basically aligns the words in source text with its counterpart in the target language using the parallel corpus provided. There are many existing word alignment algorithms that use various approaches such as statistical approach [47], [50], lexical approach [51], [52] and hybrid approach [53]. The performance of existing word alignment algorithms on Indian language has certain bottlenecks due to the language specific features such as morphological variations. The morphological variations occur on adjectives, pronouns and nouns while mapping Hindi-English language pairs. The verb morphology also affects the performance of word alignment algorithm on Indian languages. The word alignment between Hindi and Tamil was analyzed and it suffers from morphological variation in adjectives, nouns and verbs. The Table-3.1 gives examples about the need for modification in the word alignment algorithm.

Table 3.1: Morphological variations in Hindi and Tamil language

S. No.	**Morphological Variations**	**Hindi Phrase**	**English Phrase**	**Tamil Phrase**
1	Adjective Variations	**पीला** रंग (peela rang)	Yellow color	**மஞ்சள்** நிறம் (Mañcaḷ niṟam)
		पीली बोतल	Yellow bottle	**மஞ்சள்** பாட்டில்

		(peelee botal)		(mañcaḷ pāṭṭil)
2	Pronoun Variations	**मेरा** घर (mera ghar)	My house	**என்** வீடு (eṉ vīṭu)
		मेरी किताब (meri kithaab)	My book	**என்** புத்தகம் (eṉ puttakam)
		मेरे दोस्तों (mere dosthon)	My friends	**என்** நண்பர்கள் (eṉ naṇparkaḷ)
3	Noun Variations	मैंने कार को **दो घंटे** तक चलाया। (mainne kaar ko do ghante tak chalaaya.)	I drove car for two hours.	நான் **இரண்டு மணி** நேரம் கார் ஓட்டிவிட்டேன். (Nāṉ iraṇṭu maṇi nēram kār ōṭṭiviṭṭēṉ.)
		हम **दो घंटों** में दिल्ली पहुंचे। (ham do ghanton mein dillee pahunche.)	We reached delhi in two hours.	**இரண்டு மணி** நேரத்தில் நாங்கள் டெல்லி சென்றோம். (Iraṇṭu maṇi nērattil nāṅkaḷ ṭelli ceṉṟōm.)
4	Verb Variations	तुम **खेल रहे हो** (tum khel rahe ho .)	You are playing	நீங்கள் **விளையாடுகிறீர்கள்** (Nīṅkaḷ viḷaiyāṭukiṟīrkaḷ)

		तुम **खेल रही हो** (tum khel rahi ho .)	You are playing	நீங்கள் **விளையாடுகிறீர்கள்** (nīṅkaḷ viḷaiyāṭukiṟīrkaḷ)
		मैं **खेल रहा हूँ** (meiin khel raha hoon .)	I’m playing	நான் **விளையாடுகிறேன்**. (nāṉ viḷaiyāṭukiṟēṉ.)
		मैं **खेल रही हूँ** (meiin khel rahi hoon .)	I’m playing	நான் **விளையாடுகிறேன்** (Nāṉ viḷaiyāṭukiṟēṉ)
		वह **खेलेगा** (wah khelegaa)	He will play	அவர் **விளையாடுவார்** (avar viḷaiyāṭuvār)

3.3 PREPROCESSING PHASE

Since the proposed machine translation system works at word level, there is need for tokenization of source text at sentence level as well as word level. In Hindi language, there is tense, aspect and modality (TAM) information stored in the affixes of the words. These affixes also contribute to the accuracy of machine translation. To extract the TAM information stored in the affixes, longest affix matching algorithm is used to check the matching between affixes. Levenshtein distance is used to calculate the matching score. For example, consider the word खाने (khaane) and रहेगा (rahega) mentioned in input sentence – “रात को एक रोटी कम खाने से पेट हल्का रहेगा । (raat ko ek rotee kam khaane se pet halka rahega .)”. Both these words have TAM information in it and it can be extracted using the longest affix matching algorithm. The word खाने (khaane) will be analyzed and the TAM information is found as masculine, plural verb. Similarly, the TAM information of the word रहेगा (rahega) is found to be masculine, singular verb. Once these affixes are extracted, the sequence of words along with its affixes are fed to the next phase of the translation system.

3.4 HINDI PART-OF-SPEECH TAGGER

In Hindi language, there are words which have different mapping with words in Tamil as shown in Table-3.2. This can also be referred as one-to-many mapping. For example, the word खाना (khana) can be used as a noun or as a verb depending on the context in which it is used. But, the equivalent Tamil word is different according to the part-of-speech. Similarly, the Hindi word पता (Pata) has different equivalent Tamil word according to its part-of-speech. Based on bilingual corpus analysis, it is found that the one-to-many mapping can be reduced by using the part-of-speech of the word. Thus, to extract the appropriate target word for the given word in Hindi, the need for part-of-speech tagging is necessary. So, after the pre-processing is over, a word along with its affixes are fed to the part-of-speech tagger which identifies the relationship between the words.

Table 3.2: One-to-many mapping between Hindi and Tamil language

S. No.	Hindi Word	Part-of-speech	English Equivalent	Tamil Word
1	खाना (khana)	Noun	Food	சாப்பாடு (cappadu)
2	खाना (khana)	Verb	Eat	சாப்பிட (Cāppiṭa)
3	नीला (neela)	Adjective	Blue	நீலம் (Nīlam)
4	नीला (neela)	Noun	Blue	நீல (Nīla)
5	कर (Kar)	Noun	Tax	வரி (Vari)
6	कर (Kar)	Verb	Do	செய் (Cey)

7	आम (Aam)	Noun	Mango	மாம்பழம் (Māmpal̲am)
8	आम (Aam)	Adjective	Common	பொது (Podhu)
9	पता (Pata)	Noun	Address	முகவரி (Mukavari)
10	पता (Pata)	Verb	Know	தெரியும் (Teriyum)
11	साफ (Saf)	Adverb	Clean	சுத்தமான (Cuttamān̲a)
12	साफ (Saf)	Adjective	Clear	தெளிவு (Teḷivu)

In this sense-based machine translation, a hidden markov model (HMM) based POS tagger [54] was used which is as shown in Figure-3.2. The hidden markov model-based POS tagger is a statistical approach which is used to identify the probable part-of-speech of each word in the sentence. The hidden markov model-based tagger basically finds the most probable sequence of part-of-speech for a given sentence by using the transition probability $P(T_n/T_{n-1})$ and emission probability $P(W_n/T_n)$. These transition and emission probabilities are learned by the tagger using the monolingual Hindi corpus. These probabilities are calculated using the expressions,

$$P(W_n/T_n) = \frac{count\ (W_n,T_n)}{count(T_n)} \tag{3.2}$$

$$P(T_n/T_{n-1}) = \frac{count(T_n,T_{n-1})}{count(T_{n-1})} \tag{3.3}$$

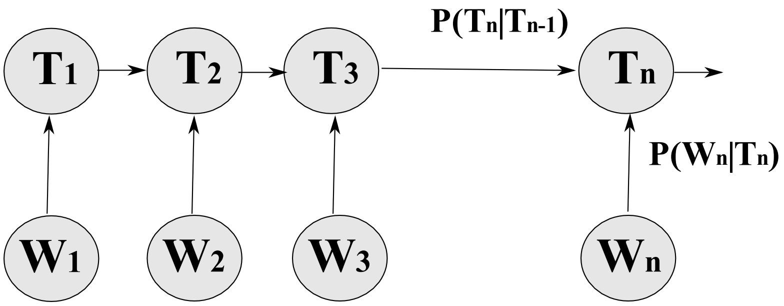

Figure 3.2: Hidden Markov model-based POS tagger

In Figure-3.2, input word sequences are denoted as W_1, W_2, W_3, … , W_n and the part-of-speech of each word is denoted as T_1, T_2, T_3, … , T_n. The part-of-speech of each word T_1, T_2, T_3, … , T_n acts as hidden states. Each of these hidden states is predicted using the emission and transition probabilities. For example, consider the input sentence as “रात को एक रोटी कम खाने से पेट हल्का रहेगा । (raat ko ek rotee kam khaane se pet halka rahega .)” and the POS tagset as [JJ (adjective), N_NN (Noun), PSP (Postposition), QT_QTC (Cardinal), QT_QTF (Quantifiers), V_VM (Verb), RD_PUNC (Punctuations), CC_CCS (Conjuncts)]. The beginning of a sentence is denoted by ^ symbol. Considering the first word “रात (raat)” in the sentence to calculate the probability of the word as a noun,

$$P(\text{रात}/N\_NN) = \frac{count\ (\text{रात},N\_NN)}{count(N\_NN)} \tag{3.4}$$

$$P(N\_NN/^\wedge) = \frac{count(N\_NN,^\wedge)}{count(^\wedge)} \tag{3.5}$$

The frequency of occurrence count is found from the monolingual Hindi corpus. Consider there are 35 occurrence of word रात (raat) as noun (N_NN) and the occurrence of noun (N_NN) to be 1000 out of which 600 times it is tagged with first word of the sentence. Number of sentences is also considered as 1000.

$$P(\text{रात}/N\_NN) = \frac{35}{1000} = 0.035 \tag{3.6}$$

$$P(N\_NN/^\wedge) = \frac{600}{1000} = 0.6 \tag{3.7}$$

The probability of word रात (raat) as any other part-of-speech is 0 since there is no occurrence of the word with different part-of-speech. Thus, the part-of-speech for word रात (raat) is found to be as N_NN. The probable part-of-speech for second word को (ko) is calculated by using the preceding word's part-of-speech as N_NN. The word को (ko) cannot have any other part-of-speech apart from postposition. The probability of को (ko) as postposition is calculated using the expression given below,

$$P\left(\text{को}/PSP\right) = \frac{count\left(\text{को},PSP\right)}{count(PSP)} \tag{3.8}$$

$$P(PSP/N\_NN) = \frac{count(PSP,N\_NN)}{count(N\_NN)} \tag{3.9}$$

Consider the occurrence of word को (ko) as postposition to be 273 and the occurrence of postposition (PSP) as 3000 for calculation. The occurrence of postposition (PSP) after a noun (N_NN) is considered as 900.

$$P\left(\text{को}/PSP\right) = \frac{273}{3000} = 0.091 \tag{3.10}$$

$$P(PSP/N\_NN) = \frac{900}{1000} = 0.9 \tag{3.11}$$

The word को (ko) is not tagged with any other part-of-speech. Thus, the word को (ko) is predicted to as postposition. Similarly, the part-of-speech of other words in the sentence are predicted and the tagged sentence will be – रात/N_NN को/PSP एक/QT_QTC रोटी/N_NN कम/QT_QTC खाने/V_VM से/PSP पेट/N_NN हल्का/QT_QTF रहेगा/V_VM ।

3.5 WORD SENSE DISAMBIGUATION

In the previous phase, the number of one-to-many mappings between Hindi and Tamil is reduced with the help the word's part-of-speech. But, it also needs further refinement in the mapping between languages. This is due to the existence of mapping between words in Hindi with the words in Tamil irrespective of the part-of-speech of the word. For example, there are words in Hindi which have multiple equivalent words in Tamil even when the part-of-speech is being used with it (as shown in Table-3.3). The Hindi word साँस (saans) being used as noun has two different Tamil words மூச்சு (muccu) and சுவாசம் (swasam). The phrase साँस द्वारा (saans dwara) will have the equivalent Tamil word for साँस as மூச்சு (muccu) whereas, the word साँस (saans) in the phrase साँस की नली (saans ki nali) will be mapped to Tamil word சுவாசம் (swasam).

Table 3.3: Necessity of word sense disambiguation in a one-to-many mapping

S. No.	Hindi Word	Part-of-speech	English Equivalent	Tamil Word
1	ताजा (taaja)	Adjective	Fresh	புதிய (Putiya)
2	ताजा (taaja)	Adjective	Fresh	சுத்தமான (sutamana)
3	साँस (saans)	Noun	Breath	மூச்சு (muccu)
4	साँस (saans)	Noun	Breath	சுவாசம் (swasam)
5	उपाय (upaay)	Noun	Remedy	மருந்துகள் (maruntukaḷ)
6	उपाय (upaay)	Noun	Solution	தீர்வு (Tīrvu)

To filter the mappings further, use of word's sense can be most appropriate option during the translation. Thus, the word sense disambiguation was introduced in this translation system. This word sense disambiguation phase makes use of the Hindi wordnet dictionary [55] to identify the various senses for the given input word. In order to narrow down to particular context in which it is being used, the identified part-of-speech is also used to retrieve the various possible senses from the wordnet. A semantic analysis between the given input words and the retrieved senses of the word provides a score of its relevance. The high scored sense is used further during the translation process. To perform the semantic analysis over the senses, a latent semantic analysis (LSA) was used. Similar kind of systems has been developed by Juan Pino et. al. [56] and Phil Katz et. al [57]. A latent semantic analysis (LSA) [58] is singular value decomposition method, which decomposes the input term-document frequency matrix (A) into three different matrices – Left singular matrix (L), Right singular matrix (R) and singular diagonal matrix (S). The left singular matrix is a singular matrix which has the eigen vectors of AA^T and it is also called as term matrix. The right singular matrix has eigen vectors of A^TA and is also known as document matrix. The singular diagonal matrix is generated using only the positive values of square roots of all non-zero eigen values of both AA^T and A^TA. The singular diagonal matrix is also known as term-by-document matrix.

$$\begin{bmatrix} f_{t1}^{1} & f_{t2}^{1} & \ldots\ldots & f_{tn}^{1} \\ f_{t1}^{2} & f_{t2}^{2} & \ldots\ldots. & f_{tn}^{2} \\ \cdots & \cdots & \cdots & \cdots \\ f_{t1}^{m} & f_{t2}^{m} & \ldots\ldots & f_{tn}^{m} \end{bmatrix} = L * R * S \tag{3.12}$$

where, f_{tn}^{m} - indicates the frequency of nth-term in mth-sentence

L - Left singular matrix (called as Term matrix)

R - Right singular matrix (called as document matrix)

S - Singular diagonal matrix (called as term-by-document matrix)

The similarity between the sentences are calculated using the cosine similarity over the dot product of the right singular matrix and singular diagonal matrix [59]. The sentences whose similarity score is found to be higher will be used further to identify the sense of the word which is under consideration. The cosine similarity is calculated using the expression mentioned below,

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i * y_i}{\left(\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}\right)} \tag{3.13}$$

Where,

x_i – denotes ith vector in right singular matrix (R)

y_i – denotes ith vector in singular diagonal matrix (S)

3.6 ALIGNMENT PHASE

Since the language model in the proposed Bayesian approach in based on the syntactic structure of the target language, there is need for alignment phase before the probability calculation that will be made in the transfer phase. Based on the analysis, it is found that the Hindi language follows a subject-object-verb (SOV) sequence which is same as in Tamil language. Hindi is a partially free word order language and Tamil is a fully free word order language. The sentence in Hindi may have multiple valid Tamil sentence with the word reorder from the original text as mentioned in the example below. Thus, there is need for grammatical restructuring of source language text with respect to the grammar of the target language text and hence, the alignment phase in the proposed system is being used before the transfer phase. Considering the following Hindi and its corresponding Tamil sentences, the examples mentioned in Table-3.4 shows the need for word alignment phase in machine translation,

Hindi Text: पानी पीना स्वास्थ्य के लिए अच्छा है।

(paanee peena svaasthy ke lie achchha hai.)

English Equivalent: Drinking water is good for health.

Tamil Text-1: குடிநீர் ஆரோக்கியத்திற்கு நல்லது.

(Kuṭinīr ārōkkiyattiṟku nallatu)

Tamil Text-2: ஆரோக்கியத்திற்கு நல்லது குடிநீர்.

(ārōkkiyattiṟku nallatu kuṭinīr)

Tamil Text-3: நல்லது ஆரோக்கியத்திற்கு குடிநீர்.

(nallatu ārōkkiyattiṟku kuṭinīr)

Table 3.4: Examples showing the need for word alignment phase

S. No.	Hindi Word	Hindi Word Position	Tamil Word	Tamil Word Position
1	पानी (paanee)	1	நீர் (nīr)	2
2	पीना (peena)	2	குடி (kuṭi)	1
3	स्वास्थ्य के लिए (svaasthy ke lie)	3	ஆரோக்கியத்திற்கு (ārōkkiyattiṟku)	3
4	अच्छा है। (achchha hai.)	4	நல்லது. (nallatu)	4

This is achieved by using statistical based alignment algorithm which requires an alignment table that provides information about the probable position of the target word. The alignment table that is being used in this phase will identify the probable position of the target word based on the source word position, source sentence length and target sentence length. The information about all these parameters is being extracted from the parallel corpus used in this approach. To generate this alignment table, the MOSES tool is being used and this tool makes use GIZA++ (a word alignment tool). GIZA++ is a free available tool for word alignment and it is provided by IBM. A sample alignment table generated from the corpus is as shown in Table-3.5 which has word position in target language (j), source word position (i), length of Hindi sentence (l) and length of its corresponding Tamil sentence (m). Consider the first entry on the Table-3.5 which uses source sentence having 8 words and target sentence with 10 words. It shows that the frequency of target word’s occurrence at position 4 provided the source word’s position is at 3 is 252. Word alignment table is generated using the positions of the words in sentence.

Table 3.5: Sample word alignment table

S.No	J (Word's Position in target language)	I (Word's Position in Source language)	L (Source Sentence Length)	M (Target Sentence Length)	Frequency of occurrence
1	4	3	8	10	252
2	7	7	10	10	231
3	1	1	6	8	115
4	3	5	7	10	152

3.7 TRANSFER PHASE

The statistical model of machine translation is designed to predict the probable target language word given its source language word. But this approach does not provide translation accuracy as the relationship between words in source and target language is one-to-many, many-to-one and many-to-many, instead of one-to-one. To improve the accuracy further, there is need for distinguishing the relationship based on the part-of-speech and also based on the word's sense.

As shown in Table-3.2, the same Hindi word with different part-of-speech has different translation in Tamil. Thus, the part-of-speech will be helpful to identify the probable target word in Tamil language. The part-of-speech of the source word is used to categorize the relationship and there is also possibility of multiple senses for word that has same part-of-speech as shown in Table-3.3. To disambiguate the appropriate sense of the word mentioned in source text, a word sense disambiguation is used. The identified senses are used in the proposed statistical machine translation approach. The transfer phase basically predicts the probable target language word based on the source language word and its part-of-speech. Using Bayes rule, the transfer phase is mathematically expressed as below,

$$P(W_t/(W_s,pos)) = P((W_s,pos)/W_t) * P(W_t) \tag{3.14}$$

Where,

W_s – Source word

pos – Part-of-speech of source word

W_t – Target Word

The source word and its part-of-speech are conditionally independent with respect to the target word. Using this assumption and by applying probabilistic rules the above expression is rewritten as,

$$P((W_s, pos)/W_t) = P(W_s/W_t) * P(pos/W_t) \tag{3.15}$$

$$P(W_s/W_t) = \frac{count(W_s,W_t)}{count(W_t)} \tag{3.16}$$

$$P(pos/W_t) = \frac{count(pos,W_t)}{count(W_t)} \tag{3.17}$$

The target word (W_t) is conditionally dependent on its previous word (W_{prev}) in the text. Similarly, every target word has dependency with its preceding word. So, the probability of target word [60] is defined as a n-gram model and is mathematically expressed as,

$$P(W_t) = P\left(W_t/W_{prev}\right) \tag{3.18}$$

$$P(W_t) = \frac{count\left(W_t,W_{prev}\right)}{count\left(W_{prev}\right)} \tag{3.19}$$

3.8 RESULTS AND DISCUSSION

For the analysis taking an example from our published word [61], consider the input text as: "ताजा साँसें और चमचमाते दाँत आपके व्यक्तित्व को निखारते हैं।" (**English Equivalent:** Fresh breath and shining teeth enhance your personality.)

After performing the initial preprocessing phase, the words in the input sentence are identified as unique token and these identified tokens are further used to extract the syntactic information in the sentence. The syntactic information is extracted using HMM based part-of-speech tagger. The sample part-of-speech tags used in this input sentence is described in the Table-3.6.

Table 3.6: Sample Part-of-speech tag-set for Hindi

S. No.	Tag	Part-of-speech
1	JJ	Adjective
2	N_NN	Noun
3	CC_CCD	Conjunction
4	PR_PRP	Pronoun
5	PSP	Postposition
6	V_VM	Finite Verb
7	V_VAUX	Auxiliary Verb
8	RD_PUNC	Punctuation

The tagged output for the text considered will be as below,

ताजा\JJ साँसें\N_NN और\CC_CCD चमचमाते\JJ दाँत\N_NN आपके\PR_PRP व्यक्तित्व\N_NN को\PSP निखारते\V_VM हैं\V_VAUX ।.\RD_PUNC.

(taaja\JJ saansen\N_NN aur\CC_CCD chamachamaate\JJ daant\N_NN aapake\PR_PRP vyaktitv\N_NN ko\PSP nikhaarate\V_VM hain\V_VAUX ।.\RD_PUNC.)

This tagged text is fed to the word sense disambiguation phase to identify the various possible senses for the word used in that context. Considering the word ताजा\JJ (taaja\JJ), the various senses for this word is retrieved from the Hindi wordnet. The various senses retrieved are – ताज़ा (taaza), ताजा (taaja), अमल्ान (amalaan), अशषुक् (ashashuk), आला (aala), गरमागरम (garamaagaram), टटका (tataka). For all the retrieved senses, its corresponding usage sentences are retrieved which are used in the word sense disambiguation phase to identify the senses that match with the input text. The identified possible senses for the word ताजा are ताज़ा and ताजा.

Using the transfer model and language model, the identified senses of the word are used to predict its probable target word. From the corpus, the frequency of word occurrence is found to generate the transfer table. Sample transfer table is shown in Table-3.7 that will be used for the illustration purpose.

Table 3.7: Sample Transfer table

S. No.	**Hindi Word**	**Tamil Word**	**Frequency of occurrence**		
			Tamil Word	**Parallel Hindi-Tamil word**	**Tamil word as adjective**
1	ताजा (taaja)	புத்துணர்ச்சியான (Puthunarchiyana)	10	8	10
2	ताज़ा (taaza)	புதுப்பித்து (Puthupithu)	8	5	2

Since there are two possible senses for the word ताजा, there are two cases to be considered,

Case-1: When the source word W_s is ताजा (taaja) with the part-of-speech tag as JJ and target word W_t as "புத்துணர்ச்சியான" (Puthunarchiyana means Fresh), the translation probability $P((W_s, pos)/W_t)$ is calculated as below,

$$P((W_s, pos)/W_t) = P(W_s/W_t) * P(pos/W_t) \tag{3.20}$$

$$P(W_s/W_t) = \frac{count(W_s, W_t)}{count(W_t)} = 8/10 \tag{3.21}$$

$$P(pos/W_t) = \frac{count(pos, W_t)}{count(W_t)} = 10/10 \tag{3.22}$$

Probability of the language model $P(W_t) = \frac{count(W_t, W_{prev})}{count(W_{prev})}$=8/30.

The overall probability $P(W_t/(W_s, pos))$ = (8/10)*(8/30) = 0.2133

Case-2: When the source word W_s is ताज़ा with the part-of-speech tag as JJ (adjective) and target word W_t as “புதுப்பித்து”(Puthupithu means New), the translation probability is calculated as below,

$P((W_s, pos)/W_t)$ = (5/8)*(2/8) (3.23)

The probability of the language model $P(W_t)$ = 1/10.

Thus, the overall probability $P(W_t/(W_s, pos))$ = (10/64)*(1/10) = 0.0156

From the above two cases, the probable translation of the word ताजा is புத்துணர்ச்சியான that has the higher probability as compared with the other one. Similarly, the words in the input sentence are translated and the generated output will be

புத்துணர்ச்சியான\JJ சுவாசம்\N_NN மற்றும்\CC_CSD பளபளப்பான\JJ பற்கள்\N_NN தங்களின்\PR_PRP தோற்றத்தை\N_NN நிர்ணயிக்கிறது\V_VM_VF .\RD_PUNC

(Puttuṇarcciyāṉa\JJ cuvācam\N_NN maṟṟum\CC_CSD paḷapaḷappāṉa\JJ paṟkaḷ\N_NN taṅkaḷiṉ\PR_PRP tōṟṟattai\N_NN nirṇayikkiṟatu\V_VM_VF .\RD_PUNC)

The Bilingual Evaluation Understudy (BLEU) score is calculated to be 0.68. The precision and recall of the system is found to be 0.85 and 0.7 respectively. As shown in Table 3.8, the system is found to be efficient as compared with the system without word sense disambiguation.

Table 3.8: Comparison of SMT (with WSD) Vs SMT (without WSD)

S. No.	Corpus Size (in number of words)	Statistical Machine Translation (with WSD)		Statistical Machine Translation (without WSD)	
		Precision (in %)	Recall (in %)	Precision (in %)	Recall (in %)
1	10000	65	63	58	57
2	20000	76	72.5	71	72
3	30000	87	86.5	83	79

3.9 CONCLUDING REMARKS

The word sense-based Hindi to Tamil machine translation system is found to perform better as compared with the system which does not use the word’s sense during translation process. The performance of HMM based part-of-speech tagger is poor when applied on the corpus provided. Thus, the performance of the word sense-based Hindi to Tamil machine translation system can be improved by using an improved Hindi part-of-speech tagger instead of the HMM based part-of-speech tagger. The proposed translation system makes use of syntactic structure and thus, its accuracy also depends on the accuracy of part-of-speech tagger which provides the syntactic information. Since the accuracy of statistical machine translation approach has dependency on the corpus size, the accuracy can be improved further with increase in the corpus size. There is restriction of the availability of corpus due to the low resource availability on both Hindi and Tamil. Thus, there is need for a better part-of-speech tagger and also an approach to handle low resource issues. The next chapter details about one such approach which is called as pivot-based approach. It also discusses about the part-of-speech tagger that performs comparatively better over HMM based part-of-speech tagger.

CHAPTER 4

PIVOT BASED MACHINE TRANSLATION SYSTEM

4.1 INTRODUCTION

In this chapter, a pivot-based statistical approach to handle low resource availability issue in the languages is under consideration. The proposed pivot-based approach also has multilayer perceptron-based part-of-speech tagger to improve the accuracy of tagging process. In case of Hindi and Tamil languages, the resource availability is poor as compared with the available resources in English language. To overcome the low-resource availability issue, English is being used as a pivot language between the source and target languages so that accuracy of the system may be improved. Pivot language-based machine translation has three major phases –

i. Source language analysis
ii. Lexical transfer (source to pivot followed by pivot to target)
iii. Structural transfer

The complete translation process is divided into two sub tasks – source to pivot language translation and then pivot to target language translation. This pivot language basically bridges the resource shortage of the two languages. The pivot language has to capture the features in the source language which can be used to generate the translation in target language. This approach also uses the syntactic feature of the source language during the translation process. The syntactic information is retrieved using a multilayer perceptron-based part-of-speech tagger for Hindi, which is further used in the transfer phase. In a pivot-based approach, the pivot (English) language text has a conditional dependency on the Hindi text and its part-of-speech. Similarly, the Tamil text has a conditional dependency on the generated English text. As per earlier discussion, the syntactic information plays a major role in improving overall performance of translation system. Therefore, part-of-speech is also considered while translating it to target language. The proposed pivot-based approach which uses the syntactic features is mathematically expressed as,

$$\boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_s},\boldsymbol{pos}}\right) = \boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_s},\boldsymbol{pos}}\right) * \boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_p},\boldsymbol{pos}}\right) \tag{4.1}$$

Where,

w_s – Source language word

w_p – Pivot language word

w_t – Target language word

pos – Part-of-speech of source word

The phases of the pivot-based Hindi to Tamil machine translation system are as shown in Figure 4.1. The input sentence is initially pre-processed to analyze both at word-level as well as at sentence-level. The word-level analysis is performed using morphological analysis, which extracts the tense, aspect and modality (TAM) information from the given sentence. The sentence is then fed to a part-of-speech tagger to extract the syntactic information and this information is fed to the transfer phase later on. Since the transfer phase has a dependency over the syntactic information, the accuracy of part-of-speech tagger will have an impact on the accuracy of translation on the whole. Thus, to improve the accuracy of part-of-speech tagger, a multilayered perceptron-based part-of-speech tagger [62] is used in this approach.

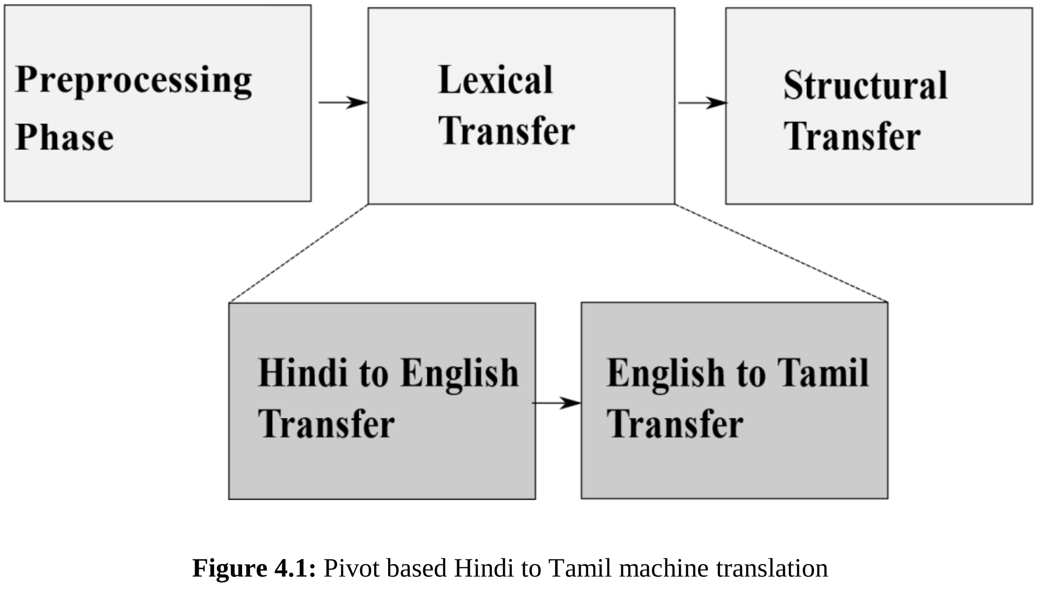

Figure 4.1: Pivot based Hindi to Tamil machine translation

4.2 PRELIMINARY PHASE

In the pivot-based translation approach, there is need to choose a pivot language which can aid in the translation process. As mentioned by M Paul et. al. [7], the pivot language can be chosen based on following parameters,

i. Resources for the pivot language should be abundant
ii. Pivot language should be related to both the source and target language
iii. Reordering properties of the language should be matching
iv. The vocabulary of the pivot language being used should be huge
v. Language family is also a major factor to select the pivot language

Since there is abundant resource on English when compared to Hindi and Tamil, English language can aid in the pivot-based translation process. As mentioned before, there is need for translation model and language model to develop a statistical machine translation system. Now, the translation model will be designed for two language pairs - source-pivot language pair and pivot-target language pair. Similarly, the language model will be developed for both pivot language and target language. The mapping between these three languages is slightly more complicated since, the English language is fixed word order language and the other two languages are free word order languages. The grammar of English, Hindi and Tamil languages also vary a lot. English language follows subject-verb-object (SVO) form, whereas, the Indian languages under consideration follows subject-object-verb (SOV) form. Since there are changes on grammar of these languages, the mapping between Hindi-English pair and English-Tamil pair can be performed using syntactic information of the languages.

Syntactic information of the languages can be retrieved by using an efficient part-of-speech tagger. To design a transfer phase of this approach, the word alignment information is more important. The basic IBM model does not consider the part-of-speech of the words during alignment process. In order to perform word alignment between these three languages, there is need for a modification on IBM model to incorporate the lexical features. This modified IBM model performs the alignment based on the part-of-speech of the word pairs. As per IBM models, the probability of words position in the target language (j) is dependent on the position of words in source language (i), length of source (l) and length of target text (m). In this modified IBM model, the part-of-speech (pos) of the source text is also considered as a parameter to predict the probable alignment for a word. Thus, the probable position of target word can be calculated using its conditional dependence with position of source word, length of source text, length of target text and part-of-speech of source word. Using Bayes theorem, the modified word alignment model is mathematically represented in below equation,

$$P\left(\frac{j}{i,l,m,pos}\right) = \frac{P(j,i,l,m,pos)}{P(i,l,m,pos)} \tag{4.2}$$

Where,

j – position of target word

i – position of source word

l – length of source text

m – length of target text

pos – part-of-speech

Using probabilistic rules and by applying Bayes theorem, the above equation can be rewritten as,

$$P\left(\frac{j}{i,l,m,pos}\right) = \frac{P\left(\frac{i}{j,l,m,pos}\right)*P(j,l,m,pos)}{P(i,l,m,pos)} \tag{4.3}$$

$$= \frac{P\left(\frac{i}{j,l,m,pos}\right)*P\left(\frac{l}{j,m,pos}\right)*P(j,m,pos)}{P(i,l,m,pos)} \tag{4.4}$$

$$= \frac{P\left(\frac{i}{j,l,m,pos}\right)*P\left(\frac{l}{j,m,pos}\right)*P\left(\frac{m}{j,pos}\right)*P(j,pos)}{P(i,l,m,pos)} \tag{4.5}$$

$$= \frac{P\left(\frac{i}{j,l,m,pos}\right)*P\left(\frac{l}{j,m,pos}\right)*P\left(\frac{m}{j,pos}\right)*P\left(\frac{pos}{j}\right)*P(j)}{P(i,l,m,pos)} \tag{4.6}$$

The length of source and target text are independent events with respect to the position of target word and part-of-speech of source word. This is due to the free word order feature of both Hindi and Tamil language. Based on this criterion, the above equation is reduced to,

$$P\left(\frac{j}{i,l,m,pos}\right) = \frac{P\left(\frac{i}{j,l,m,pos}\right)*P(l)*P(m)*P(pos)*P(j)}{P(i,l,m,pos)} \tag{4.7}$$

$$P(i,l,m,pos) = P\left(\frac{pos}{i,l,m}\right) * P(i,l,m) \tag{4.8}$$

$$= P\left(\frac{pos}{i,l,m}\right) * P\left(\frac{m}{i,l}\right) * P(i,l) \tag{4.9}$$

$$= P\left(\frac{pos}{i,l,m}\right) * P\left(\frac{m}{i,l}\right) * P\left(\frac{l}{i}\right) * P(i) \tag{4.10}$$

The part-of-speech of source word and target sentence length are two independent events. The length of target text has no dependency with the position of source word. Similarly, the length of source sentence does not have dependency with position of source word. Thus, by neglecting these independent events, this equation is further reduced to,

$$P(i,l,m,pos) = P\left(\frac{pos}{i,l}\right) * P\left(\frac{m}{l}\right) * P(l) * P(i) \tag{4.11}$$

$$P\left(\frac{j}{i,l,m,pos}\right) = \frac{P\left(\frac{i}{j,l,m,pos}\right)*P(l)*P(m)*P(pos)*P(j)}{P\left(\frac{pos}{i,l}\right)*P\left(\frac{m}{l}\right)*P(l)*P(i)} \tag{4.12}$$

$$P\left(\frac{j}{i,l,m,pos}\right) = \frac{P\left(\frac{i}{l}\right)*P(m)*P(pos)*P(j)}{P\left(\frac{pos}{i,l}\right)*P\left(\frac{m}{l}\right)*P(i)} \tag{4.13}$$

The probabilities on the right-hand side of the above equation can be calculated based on the corpus that is being used for training purpose. For generating the alignment table and translation table, mgiza++ tool is used which is a part of the MOSES tool developed by Koehn et al [30].

4.3 PREPROCESSING PHASE

The syntactic information of the input text is identified using a part-of-speech tagger. Before applying the part-of-speech tagging process, there is need for word-level analysis which is performed using a Hindi morphological analyzer. The Hindi morphological analyzer basically makes use of a longest affix matching algorithm which in turn uses the Levenshtein distance measure to extract the morphology of the words in input text. Once the word’s morphology has been extracted, it is fed to a part-of-speech tagger to identify the syntactic information. The Hindi part-of-speech tagger is developed using a multilayer perceptron-based neural network [62]. A multilayer perceptron based neural network is defined as a feedforward neural network that has one or more hidden layers in between input layer and output layer. It is generally used to classify non-linear data. It also uses supervised learning with the help of backpropagation algorithm. For training this neural network, the features identified are - probability of the word at position *'i'* given the word's tag $P(word_i/tag_i)$ and the probability of the current word's tag at position *'i'* given the previous word's tag at position *'i-1'*, $P(tag_i/tag_{i-1})$. Based on the tagged corpus, these probability combinations are found and stored in a vector. This vector is fed to the input layer of multilayer perceptron for the training purpose. During training, each of these vectors is used to predict the optimum weights for the network. The weights are changed based on the error between predicted output and actual output. This is known as supervised learning. Supervised learning in this network is performed through backpropagation algorithm. The multilayer perceptron-based part-of-speech tagger is shown in Figure-4.2,

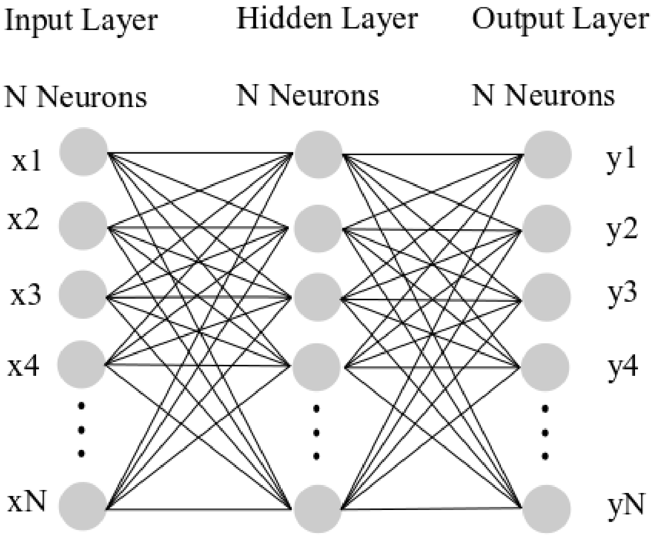

Figure 4.2: Multilayer Perceptron based POS Tagger

The hidden layer in the multilayered perceptron tagger is activated using a sigmoid function [63] which is expressed mathematically as below,

$$y(f(x)) = \frac{1}{1+e^{-f(x)}} \quad (4.14)$$

$$f(x) = \sum_{j=1}^{N} w_j * P\left(\frac{word}{tag_j}\right) * P\left(\frac{tag_j}{tag_i}\right) \qquad (4.15)$$

where,

$word$ – current word being processed

tag_i – Part-of-speech tag of previous word

tag_j – Part-of-speech tag from the considered POS tag set

N – number of distinct part-of-speech tags

Using the multilayer perceptron-based part-of-speech tagger, there will be an output matrix $[y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_N]$ where the dimension 'N' is equivalent to the number of distinct tags. The maximum value from this output matrix is used to locate the probable tag of word. For illustration, consider the current word as "खाने (khaane)" in the phrase "रात को एक रोटी कम खाने से पेट हल्का रहेगा I (raat ko ek rotee kam khaane se pet halka rahega .)" and the POS tag set as [JJ (adjective), N_NN (Noun), PSP (Postposition), QT_QTC (Cardinal), QT_QTF (Quantifiers), V_VM (Verb), RD_PUNC (Punctuations), CC_CCS (Conjuncts)]. Assume, the frequency of the word खाने in the corpus is 500 and in that it occurs as noun for 200 times. The rest of the occurrence is assumed as a verb. Previous word's part-of-speech is cardinal (QT_QTC) and its occurrence before a noun or a verb is equal i.e., 250 times each. The occurrence of cardinal (QT_QTC) in the corpus is 1000 times. The probability of "खाने (khaane)" as a noun is,

$$P\left(\frac{\text{खाने}}{N\_NN}\right) = \frac{count(\text{खाने}, N\_NN)}{count(N\_NN)} = \frac{200}{500} = 0.4 \qquad (4.16)$$

$$P\left(\frac{N\_NN}{QT\_QTC}\right) = \frac{count(N\_NN, QT\_QTC)}{count(QT\_QTC)} = \frac{250}{1000} = 0.25 \qquad (4.17)$$

Similarly, the calculation for verb is calculated and the probability values are,

$$P\left(\frac{\text{खाने}}{N_{NN}}\right) = \frac{300}{500} = 0.6 \qquad (4.18)$$

$$P\left(\frac{N\_NN}{QT\_QTC}\right) = \frac{250}{1000} = 0.25 \qquad (4.19)$$

Using the probability calculated in equation (4.16) to equation (4.19), the input vector is formed as [0, 0.1, 0, 0, 0, 0.15, 0, 0]. The vector values are placed in the order of the tag set being used. This input vector is fed to the multilayer perceptron based neural network and it generates the output vector with N values. The value in output vector which is maximum is considered and mapped with the tag set to identify the probable part-of-speech.

4.4 LEXICAL TRANSFER PHASE

The syntactic feature identified using the multilayer perceptron-based part-of-speech tagger is fed to the next phase, called as lexical transfer phase. Since there is an intermediate pivot language, the lexical transfer phase works in two stages translating Hindi to English and then English to Tamil. In the transfer phase-I, the Hindi language is translated to English using Bayesian approach. The translated English sentence from transfer phase-I is fed to transfer phase-II which in turn translates it to Tamil language. Pivot language word has a conditional dependency with source word and its part-of-speech. Thus, by considering the prior knowledge about the source word and its part-of-speech, the probability of pivot word will be calculated using the probability $\boldsymbol{P}\left(\frac{w_p}{w_s,pos}\right)$. Similarly, the target language word is conditionally dependent on identified pivot word and the part-of-speech. Using this condition, the probable target word will be predicted by using the probability $\boldsymbol{P}\left(\frac{w_t}{w_p,pos}\right)$. The transfer phase is defined in terms of these two probabilities and is mathematically represented as,

$$\boldsymbol{P}\left(\frac{w_t}{w_s,pos}\right) = \boldsymbol{P}\left(\frac{w_p}{w_s,pos}\right) * \boldsymbol{P}\left(\frac{w_t}{w_p,pos}\right) \tag{4.20}$$

4.4.1. TRANSFER PHASE-I (SOURCE TO PIVOT)

Since there is translation from source (Hindi) to pivot (English) language, there is need for mapping between these two languages. The Hindi language is basically free word order language, whereas the English language is a fixed word order language. Thus, the complexity of mapping increases to various types such as one-to-one, one-to-many, many-to-one and many-to-many. Few examples of different types of mapping between Hindi and English is mentioned in Table-4.1. The examples “paani” and “kaam” mentioned in the table below is an example of one-to-one mapping. The equivalent English words for these two Hindi words is “water” and “work”. During translation, there is a direct one-to-one mapping between these words and there is no other possible English word that can replace it. The one-to-one mapping is direct word-to-word mapping thus reducing the complicacy in it.

Table 4.1: Types of mapping between Hindi and English

S. No.	Hindi Word/Phrase	English Word/Phrase	Type of mapping
1	पानी (paani)	Water	One-to-one
2	काम (kaam)	Work	One-to-one
3	पूजा (pooja)	Worship or Pooja	One-to-many
4	सोना (sona)	Gold or Sleep	One-to-many
3	राम ने (raam ne)	Ram	Many-to-one
4	आराम से (aaraam se)	Comfortably	Many-to-one
5	कर रहा है (kar raha hai)	Is doing	Many-to-many
6	माना जाता (maana jaata)	Is believed	Many-to-many

In the phrases “ram ne” and “aaraam se”, there is a many-to-one mapping and these words will be mapped with the English words “ram” and “comfortably”. The phrases “kar raha hain” and “maana jaata” has many-to-many mapping where multiple words occur in sequence. It will be mapped with the corresponding multiple English word sequence. For both many-to-one and many-to-many mapping, there is need for statistical information about the words that can be grouped together in sequence to assist the translation process. For this purpose, the system makes use of n-gram statistical analysis where 'n' consecutive source words are considered during the transfer phase.

"Pooja" and "sona" are two sample words that have multiple English target words and thus are examples of one-to-many mapping. This is a special case of mapping where the multiple words in target do not occur together and not in the same sentence too. The word "sona" will be mapped to the word "gold" when it is used as noun. The same word will be mapped to "sleep" when it acts as verb. Thus, the one-to-many mapping can be handled by using the syntactic information captured in the previous phase. Based on these mappings, the statistical information is extracted from the parallel bilingual corpus. This information is used to develop the statistical transfer phase between source and pivot language. Statistical transfer phase-I is represented mathematically as follows,

$$\boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_s, pos}}\right) = \boldsymbol{P}\left(\frac{\boldsymbol{w_s, pos}}{\boldsymbol{w_p}}\right) * \boldsymbol{P}(\boldsymbol{w_p}) \tag{4.21}$$

Where,

w_s – Source language word

w_p – Pivot language word

pos – Part-of-speech of source word

As discussed before, the word "sona" has different mapping according to its part-of-speech. Thus, the probability $\boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_s, pos}}\right)$ is used to find the probable pivot word based on source word and its part-of-speech. The probable pivot word will be "gold" for Hindi word "sona" if it is tagged as noun, whereas, it will be "sleep" if the word "sona" is tagged as verb.

Since the source word and its part-of-speech does not depend on the pivot language word, the above expression (4.21) is reduced to,

$$\boldsymbol{P}\left(\frac{\boldsymbol{w_s,pos}}{\boldsymbol{w_p}}\right) = \boldsymbol{P}\left(\frac{\boldsymbol{w_s}}{\boldsymbol{w_p}}\right) * \boldsymbol{P}\left(\frac{\boldsymbol{pos}}{\boldsymbol{w_p}}\right) \tag{4.22}$$

$$\boldsymbol{P}(\boldsymbol{w_p}) = \boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_{p-1}}}\right) \tag{4.23}$$

4.4.2. TRANSFER PHASE-II (PIVOT TO TARGET)

This phase also works in similar manner as it was done in previous source-pivot transfer phase. The target language (i.e., Tamil) is fully free word order language which is in contrast to the English language and both the languages follow different grammatical rules. In Tamil language, words are formed by stacking multiple suffixes and/or prefixes onto the root word. Thus, the Tamil language is also called as agglutinative language [64]. Due to agglutination, all the types of mapping except many-to-many is applicable for the English and Tamil language. Sample mappings between English and Tamil language are as shown in Table-4.2.

Table 4.2: Types of mapping between English and Tamil

S. No.	**English Word/Phrase**	**Tamil Word/Phrase**	**Type of mapping**
1	Water	நீர் (neer)	One-to-one
2	Work	வேலை (velai)	One-to-one
3	Book	பதிவு/புத்தகம் (pativu/puttakam)	One-to-many
4	Remedy	மருந்து/தீர்வு (marunthu/teervu)	One-to-many
5	Is doing	செய்கிறார் (ceykiṟār)	Many-to-one

6	The medicine	மருந்து (Maruntu)	Many-to-one

The one-to-many mapping between English and Tamil can be disambiguated with the help of syntactic information. The many-to-one mapping can be used in translation with the help of n-gram approach. In this approach ‘n’ pivot language words are used to find the most probable translation for it. The word alignment model is once again needed in this phase of translation to assist in the statistical approach. Thus, there will be generation of alignment table as well as a translation table using MGIZA++ tool with a slight modification on the table. In general, the alignment table has the position of target word (j), position of pivot word (i), length of pivot text (l), length of target text (m) and probability (p(j/i, l, m)). The alignment table in this proposed system has been modified by including part-of-speech of the pivot word in addition to all other parameters. The sample alignment table is shown in Table-4.3. The transfer table has the pivot word (w_p), part-of-speech of the pivot word (pos), its equivalent target word (w_t) and their respective frequency of occurrence. The sample transfer table is as shown in Table-4.4.

Table 4.3: Sample Alignment table for English-Tamil language pair

S.No	Part-of-speech	J (Word’s Position in target language)	I (Word’s Position in Source language)	L (Source Sentence Length)	M (Target Sentence Length)	Frequency of occurrence
1	Noun	1	1	6	4	850
2	Noun	2	9	9	5	734
3	Verb	4	3	6	5	354

Using extended Bayes theorem, the transfer phase-II can be represented mathematically as,

$$\boldsymbol{P}\left(\frac{w_t}{w_p,pos}\right) = \frac{P\left(\frac{w_p,pos}{w_t}\right)*P(w_t)}{P(w_p,pos)} \tag{4.24}$$

The probability $\boldsymbol{P(w_p, pos) \cong 1}$ due to the direct mapping between part-of-speech and pivot word. Thus, the expression (4.24) is rewritten as,

$$\boldsymbol{P\left(\frac{w_t}{w_p, pos}\right) = P\left(\frac{w_p, pos}{w_t}\right) * P(w_t)} \tag{4.25}$$

The part-of-speech and pivot word are independent events with reference to target word and thus, the above expression (4.25) is modified as,

$$\boldsymbol{P\left(\frac{w_t}{w_p, pos}\right) = P\left(\frac{w_p}{w_t}\right) * P\left(\frac{pos}{w_t}\right) * P\left(\frac{w_t}{w_{t-1}}\right)} \tag{4.26}$$

Table 4.4: Sample transfer table for English-Tamil language pair

S. No.	English Word	Tamil Word	Part-of-speech	Frequency of occurrence
1	Read	படிக்க (Paṭikka)	Verb	135
2	Heartbeat	இதயத்துடிப்பு (Itayattuṭippu)	Noun	512

4.5 STRUCTURAL TRANSFER

The sentence that is generated from the lexical transfer phase will not be grammatically correct with reference to the target language. In case of Hindi and Tamil language, both of these follow subject-object-verb (SOV) form. But, the Hindi language is partially free word ordered as compared with Tamil which is fully free word order language. This increases the necessity for the structural transfer phase. The structural transfer in the proposed system makes use of naïve Bayes model [65] that is developed using the alignment table generated by the GIZA++ tool. Naïve Bayes model designed by using the positional information from the source text, length of sentences in both languages and the part-of-speech of the words. The grammatical correctness is obtained by taking the probable sequence of part-of-speech which has the maximum probability as per the naïve Bayes approach. The target text is rearranged according to the probable sequence that is identified. The position of source word has influences over the target word’s position. Thus, the alignment table constructed has the information about the probable target word’s position provided the source word’s position is mentioned. The source word’s position is extracted from the input text itself.

4.6 RESULTS AND DISCUSSION

This proposed system has been developed using the parallel corpus on three languages and each monolingual corpus having around 25000 sentences in it. This parallel corpus has been provided by Technology Development for Indian Languages (TDIL) programme initiated by the Department of Electronics and Information Technology (DeitY), Ministry of Communication and Information Technology (MC&IT), Govt. of India. The system has been evaluated with varying set of data to analyze its performance. An example below illustrates the working of various phases in this pivot-based statistical approach,

Hindi text: एक सामान्य व्यक्ति का ब्लड प्रेशर 140 - 90 से कम होना चाहिए ।

(ek saamaany vyakti ka blad preshar 140 - 90 se kam hona chaahie .)

This source text is subjected to preprocessing in which the sentence is tokenized into words and the words are analyzed using morphological analysis. The words are further fed to part-of-speech tagger which was developed in python 3.5. The parameters of this multi-layer perceptron-based POS tagger was learnt until the error is as minimum as possible. The final network used after training has a mean square error of 0.06 after 32 epochs. The output from the tagger for the sentence considered is as below,

एक/QT_QTC सामान्य/JJ व्यक्ति/N_NN का/PSP ब्लड/N_NN प्रेशर/N_NN 140/QT_QTC -/RD_SYM 90/QT_QTC से/PSP कम/QT_QTF होना/V_VAUX चाहिए/V_VAUX ।/RD_PUNC

The tagged text is further fed to transfer phase-I which translates the text to its probable English text. Let us consider the word सामान्य/JJ (saamaany) for the discussion. The word सामान्य/JJ has two entries on the transfer table as shown below,

Table 4.5: Transfer table for the Hindi word “saamaany”

S. No.	Hindi Word	Part-of-speech Tag	English Word	Frequency of occurrence
1	सामान्य	JJ (Adjective)	General	5
2	सामान्य	JJ (Adjective)	Normal	11

The probability $\boldsymbol{P}\left(\frac{w_p}{w_s, pos}\right)$ is calculated using the transfer table shown above. For instance, consider the English word (pivot) as “General” which has frequency of occurrence in the corpus as 19. The word “General” occurs as adjective and has a frequency of occurrence as 12. Assume, the probable pivot word for preceding word “एक (ek)” is “A” and its frequency of occurrence as 54. The word “General” is preceded by word “A” in 10 occurrences. The probability calculation to find the probable pivot word is as below,

$$\boldsymbol{P}\left(\frac{w_p}{w_s, pos}\right) = \boldsymbol{P}\left(\frac{w_s}{w_p}\right) * \boldsymbol{P}\left(\frac{pos}{w_p}\right) * \boldsymbol{P}\left(\frac{w_p}{w_{p-1}}\right) = \frac{5}{19} * \frac{12}{19} * \frac{10}{54} = \mathbf{0.0307} \qquad (4.27)$$

Considering the pivot word (in English) as “normal” which has frequency of occurrence in the corpus as 24. The word “Normal” is tagged as adjective in 12 of its occurrences and it is preceded by the word "A" in 2 instances.

$$\boldsymbol{P}\left(\frac{w_p}{w_s, pos}\right) = \boldsymbol{P}\left(\frac{w_s}{w_p}\right) * \boldsymbol{P}\left(\frac{pos}{w_p}\right) * \boldsymbol{P}\left(\frac{w_p}{w_{p-1}}\right) = \frac{11}{24} * \frac{12}{24} * \frac{2}{54} = \mathbf{0.0085} \qquad (4.28)$$

Thus, according to the above calculations, it is evident that the probable pivot word for the word सामान्य/JJ (saamaany) is “general”. Similarly, all the words in input sentence are translated to pivot language and the generated text is “a/DT general/JJ man/N_NN blood/N_NN pressure/N_NN 140/QT_QTC –/RD_SYM 90/QT_QTC less/QT_QTF than/PSP should/V_VAUX be/V_VAUX”.

This intermediate text is fed to transfer phase-II and the calculation for translating the word from “general” to its equivalent Tamil word is as follows,

The word “General” can be mapped with the word “பொது (podhu)” and it occurs 10 times in the Tamil corpus. Word “General” occurs in parallel with the word “பொது (podhu)” in 6 instances. The word “பொது (podhu)” is tagged as adjective in all its occurrences. Assume, the preceding probable target word is found to be “ஒரு (oru)” and it occurs 23 times in the corpus. The occurrence of word “ஒரு (oru)” before the word “பொது (podhu)” is 8. Using these frequencies, the probability calculation is as below,

$$\boldsymbol{P}\left(\frac{w_t}{w_p, pos}\right) = \boldsymbol{P}\left(\frac{w_p}{w_t}\right) * \boldsymbol{P}\left(\frac{pos}{w_t}\right) * \boldsymbol{P}\left(\frac{w_t}{w_{t-1}}\right) = \frac{6}{10} * \frac{10}{10} * \frac{8}{23} = \mathbf{0.2087} \qquad (4.29)$$

The target word is thus predicted to be “பொது (podhu)” for the English word “general”. Similarly, all other words are translated to its probable Tamil word. In this case, the grammar of the input sentence matches with the target text. Thus, there is no need for structural rearrangements. This is identified by using the naïve Bayes approach described before. The generated sentence is compared with the reference sentence and is as shown in the table below,

Table 4.6: Comparison of generated target text with its reference text

Generated Text	ஒரு (oru)	பொது (podhu)	நபரின் (naparin)	இரத்த (iratta)	அழுத்தம் (aluttam)	140 – 90	குறைக்க (kuraikka)	வேண்டும். (ventum)
Reference Text	ஒரு (oru)	சாதாரண (catarana)	நபரின் (naparin)	இரத்த (iratta)	அழுத்தம் (aluttam)	140 – 90	குறைக்க (kuraikka)	வேண்டும். (ventum)
Match	1	0	1	1	1	1	1	1

This proposed system has been analyzed on various other source sentence and its evaluation metrics are calculated. The results of evaluation metrics are as shown in Table-4.7 which details about the precision and recall of the proposed pivot language-based system. The table also shows the comparison of proposed system with the statistical machine translation without pivot language.

Table 4.7: Comparison of pivot-based statistical MT Vs simple Statistical MT (without pivot)

S. No.	**Corpus Size (in number of words)**	**Pivot Based Hindi-Tamil Statistical Machine Translation**		**Hindi to Tamil Statistical machine translation**	
		Precision (in %)	**Recall (in %)**	**Precision (in %)**	**Recall (in %)**
1	10000	53	52	58	57
2	20000	64	67	71	72
3	30000	76	74.5	83	79

The performance of the pivot-based machine translation system seems to be poor as compared with the statistical machine translation without pivot language which is evident from the graph shown in Figures 4.3 and 4.4. The BLEU score of this pivot-based system is found to be 0.54.

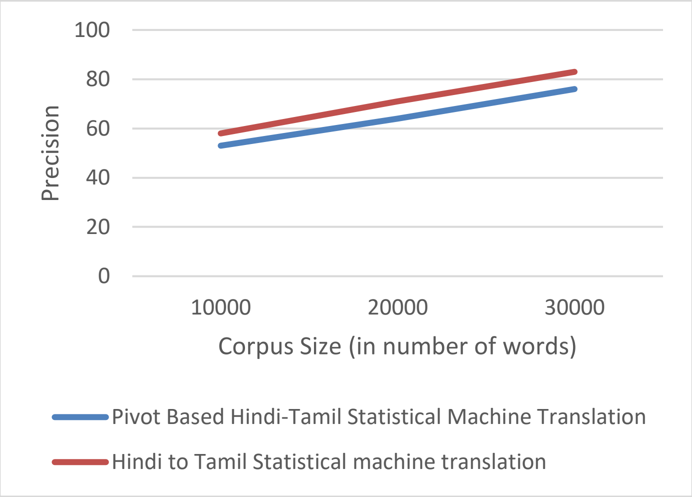

Figure 4.3: Comparison of pivot-based SMT with SMT (without pivot) in terms of precision

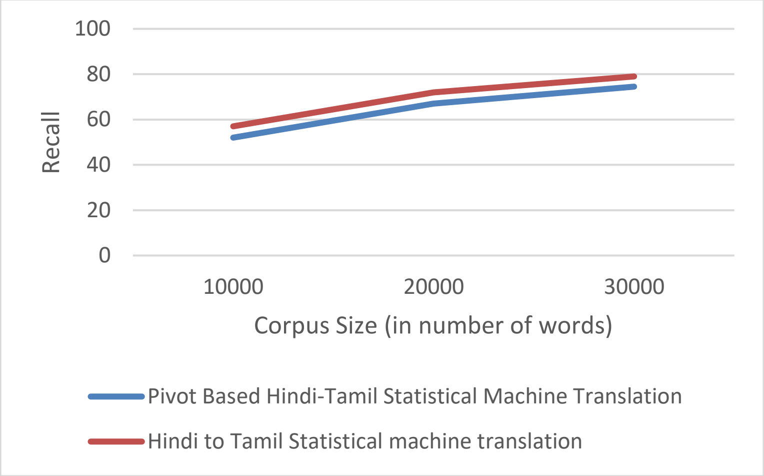

Figure 4.4: Comparison of pivot-based SMT with SMT (without pivot) in terms of recall

4.7 CONCLUDING REMARKS

Based on the comparison it is found that there is almost no improvement in accuracy in a pivot-based system. This is due to the loss of semantics during the translation process. That is, the context of the word in the input Hindi language is distorted during its translation through the pivot language, English. Apart from the semantics, there is more ambiguity introduced when there is mapping with the help of pivot language. Thus, increasing the distortion of the information stored in input text during the translation, which in turn degrades the performance of the overall system. This can be handled by introducing the semantic features during the pivot-based translation process. In the next chapter, to overcome the loss of semantics, a hybrid approach is introduced. Hybrid approach uses word sense information in a pivot-based approach.

CHAPTER 5

HYBRID APPROACH FOR HINDI TO TAMIL TRANSLATION

5.1 INTRODUCTION

In this chapter, a hybrid approach is being proposed for the translation to further reduce loss of semantics during the pivot based Hindi to Tamil machine translation, besides that provided by word sense disambiguation. The ambiguity introduced by the use of pivot language has also been reduced by the use of semantic information during the translation process. Thus, the main aim of this hybrid approach is to handle the low resource availability issue without loss of the semantic information contained in the source text. The major phases designed in this hybrid approach besides preprocessing are,

i. Source language analysis
    a. Hindi part-of-speech tagger
    b. Word sense disambiguation
ii. Lexical transfer (source to pivot followed by pivot to target)
iii. Structural transfer

As used in the pivot-based approach, there is a lexical transfer phase which occurs in two steps – Hindi to English and then English to Tamil. But it was having semantic distortion due to which the overall performance is not up to the mark. The word sense disambiguation was introduced in the source language analysis phase, so that the sense information will be made available for the lexical transfer phase. The system architecture of this hybrid approach is shown in Figure 5.1. This system considers both the syntactic and semantic features to improve the accuracy of the machine translation system. These features are extracted during the source language analysis phase and are fed to the lexical transfer phase. The lexical transfer is performed on various identified relevant senses of the words mentioned in input sentence. The lexical transfer from Hindi to English is performed before the lexical transfer from English to Tamil. Once the lexical transfer is performed, the structural rearrangements is performed in the structural transfer phase. The structural rearrangement is performed to keep the generated target sentence in a grammatically correct manner according to the target language grammar rules.

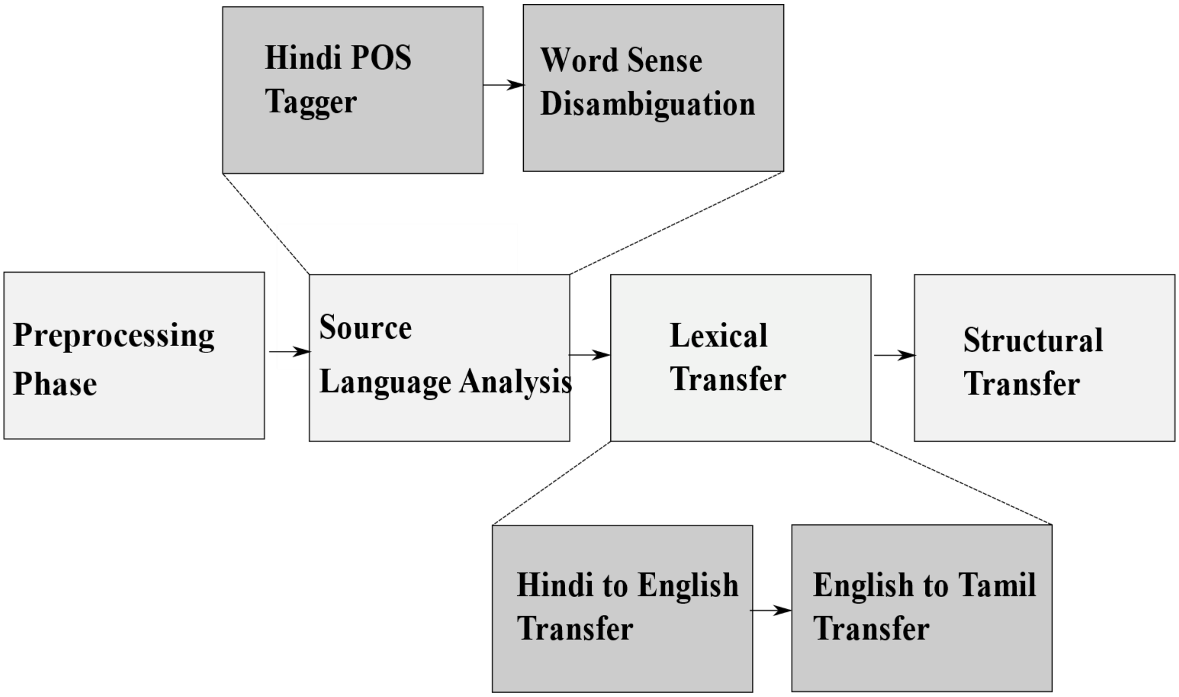

Figure 5.1: Architecture of hybrid machine translation approach

5.2 PRELIMINARY PHASE

The hybrid approach makes use of both pivot language and word sense disambiguation. The lexical transfer phase used in hybrid approach makes use of syntactic as well as semantic features. Thus, there is need for a translation model and language model which considers both the syntactic and semantic information. The translation model and language model has been developed for both language pairs – Hindi-English and English-Tamil. These translation model and language model are developed based on the parallel corpus provided by Technology Development for Indian Languages (TDIL) programme, Department of Information Technology (DIT), Ministry of Communication & IT, Government of India. The parallel corpus used has around 25000 sentence pairs and that too specifically on health domain. The lexical transfer phase can be mathematically represented as,

$$\boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_s,pos}}\right) = \boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_s,pos}}\right) * \boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_p,pos}}\right) \tag{5.1}$$

Where,

w_s – Source language word

w_p – Pivot language word

w_t – Target language word

pos – Part-of-speech of source word

Which is same as in pivot-based approach except that the $\boldsymbol{w_s}$ mentioned in above equation is over the various identified relevant senses of the word occurring in the input source text. The relevant senses are identified using the word sense disambiguation. The target word whose probability is maximum will be the most appropriate target word for the given source word.

5.3 PREPROCESSING PHASE

The input sentence is subjected to tokenization at word level using a tokenization process. Once the words are tokenized, there is need for analysis on the words to gather the information about the morphology of the words. As in earlier approaches, the morphology of the words is extracted using longest affix matching algorithm which in turn uses the Levenshtein distance to find the nearest affix that can be stripped from the input word. The affixes which are separated from the root word should not be discarded since, they have the tense, aspect and modality (TAM) information. A list of affixes along with its information are maintained for identifying the affixes in the words. Once the morphological information of the words is extracted, the syntactic and semantic relation between the words need to be analyzed. The relation between words are analyzed and its syntactic feature is extracted in source language analysis phase.

5.4 SOURCE LANGUAGE ANALYSIS

This language analysis phase is divided into two sub-modules – part-of-speech tagger module and word sense disambiguation module. The Hindi part-of-speech tagger is used to extract the syntactic information from the given text. The word sense disambiguation phase is used to extract the semantic information from the source text. The performance of proposed hybrid approach has dependency over both the extracted information.

5.4.1 HINDI PART-OF-SPEECH TAGGER

Hindi part-of-speech tagger extracts syntactic features from the source text and it is developed using a multilayer perceptron-based neural network [62]. Since, this machine translation system basically uses syntactic and semantic information, the accuracy of the tagger is more important for the accuracy of translation. Hence, the multilayer perceptron-based part-of-speech tagger was used instead of simple hidden Markov model (HMM) based part-of-speech tagger. The input and output layer size of multilayer perceptron-based tagger will be same as the number of part-of-speech tags being used in the system. Only one hidden layer is used in this multilayer perceptron-based tagger. It has been found that with more number of hidden layers the performance degrades instead of improving. The hidden layer is activated using a sigmoid function [63] which is mathematically represented as below,

$$y\big(f(x)\big) = \frac{1}{1+e^{-f(x)}} \tag{5.2}$$

$$f(x) = \sum_{j=1}^{N} w_j * P\left(\frac{word}{tag_j}\right) * P\left(\frac{tag_j}{tag_i}\right) \tag{5.3}$$

where,

$word$ – current word being processed

tag_i – Part-of-speech tag of previous word

tag_j – Part-of-speech tag from the considered POS tag set

N – number of distinct part-of-speech tags

Both probabilities mentioned in $f(x)$ are generated using the input corpus and are further fed to the multilayer perceptron-based tagger for training purposes. The various possible part-of-speech of the input word are predicted using the word's position and its preceding word's part-of-speech. Once it has been trained, the identified probable part-of-speech of a word is fed to the system as input and the system will provide a vector as an output. This vector is mapped with the part-of-speech to decode the exact part-of-speech of the given word.

5.4.2 WORD SENSE DISAMBIGUATION

Once the extraction of syntactic information is performed, the sentence is subjected to sense disambiguation phase to identify appropriate sense for the word mentioned in that context. This proposed system makes use of latent semantic analysis (LSA) to perform sense disambiguation [58]. The singular value decomposition is performed in a similar manner as mentioned in chapter-3 before. The term-frequency matrix is generated using sentences that uses the same input word but in different context. The various contextual sentences are retrieved from the Hindi wordnet [55]. The generated term-frequency matrix is decomposed to three different matrices and the similarity between sentences are calculated to identify the appropriate sense for the input word. To identify the similar sentence, cosine similarity is applied over the dot product of the right singular matrix and the singular diagonal matrix. The sentence which has maximum cosine similarity value is considered as the appropriate sense of input word mentioned. The identified syntactic and semantic information are used by the lexical transfer phase for translation purpose.

5.5 LEXICAL TRANSFER

The lexical transfer phase is used to translate the source word to its appropriate target word using the statistical information generated from the parallel bilingual corpus. Since there is an intermediate pivot language, the lexical transfer phase has to work in two steps – lexical transfer phase-I and lexical transfer phase-II. These two modules work in the same manner as mentioned in chapter-4. In addition to that, the lexical transfer phase makes use of the word sense information to reduce the number of possible ambiguous output which in turn reduces the semantic distortion that occurred in the pivot-based approach (without word sense disambiguation). The phase is mathematically expressed as,

$$\boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_s},\boldsymbol{pos}}\right) = \boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_s},\boldsymbol{pos}}\right) * \boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_p},\boldsymbol{pos}}\right) \tag{5.4}$$

Using extended Bayes theorem, the above expression (5.4) is rewritten as,

$$\boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_s},\boldsymbol{pos}}\right) = \left[\boldsymbol{P}\left(\frac{\boldsymbol{w_s}}{\boldsymbol{w_p}}\right) * \boldsymbol{P}\left(\frac{\boldsymbol{pos}}{\boldsymbol{w_p}}\right)\boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_{p-1}}}\right)\right] * \left[\boldsymbol{P}\left(\frac{\boldsymbol{w_p}}{\boldsymbol{w_t}}\right) * \boldsymbol{P}\left(\frac{\boldsymbol{pos}}{\boldsymbol{w_t}}\right) * \boldsymbol{P}\left(\frac{\boldsymbol{w_t}}{\boldsymbol{w_{t-1}}}\right)\right] \tag{5.5}$$

The hybrid approach for Hindi to Tamil machine translation is compared with the naïve Bayes statistical machine translation system in terms of the features that are being used in both the system. The comparison shown in Table 5.1 illustrates about the advantages and disadvantages in both the system. Both the proposed hybrid approach and naïve Bayes statistical approach makes use of the language model and translation model. But, in hybrid approach, a pivot language is being used to handle the low resource availability issue, due to which the semantics may get lost while translating from source to target through pivot language. In order to handle this distortion, the word sense feature was introduced in the proposed approach which does not exist in the naïve Bayes approach. Due to the word's morphological structure in Hindi and Tamil language, the part-of-speech of the word can be used as a feature to predict the appropriate translation. Thus, part-of-speech was also used in the proposed approach. The alignment model that was used in naïve Bayes approach doesn't consider the part-of-speech of the words which was introduced in the proposed method since there is free word order nature in the languages, Hindi and Tamil.

Table 5.1: Comparison of proposed machine translation with Naïve Bayes statistical machine translation

S. No.	Features	Naïve Bayes Statistical Machine Translation	Advantage/Disadvantage	Proposed Hybrid Machine Translation	Advantage/Disadvantage
1	Language Model *P(s)*	Used for source language	It considers preceding word during translation mechanism which improves the translation	Used for source language as well as pivot language	It considers preceding word during translation and pivot language is used as there is low resource availability

2	Translation Model *P(s|t)*	Used for target language	It maps the source word with the target language and predicts the probable translation which also helps during translation	Used for target as well as pivot language	It has the same advantage as in Naïve Bayes but used pivot since the languages are low resource language
3	Part-of-speech (POS)	Not used	It is a disadvantage	Used in the translation model which is modified accordingly	POS provides more accurate mapping between source and target words. Thus, it improves the translation accuracy
4	Alignment Model	Used for word alignment	Uses traditional IBM models and doesn't consider POS during alignment	Used for word alignment but has been modified to consider part-of-speech during alignment process	Uses modified IBM model and its alignment improves since the word position depends on the POS of the word as well
5	Word Sense	Since it works only on probabilistic manner, it does not include sense identification.	Since contextual information is not considered translation is poor in certain cases	During translation phase, it considers the words sense in the context it is being used	Improves the translation based on contextual information. Words translation differs based on the sense it is being used

5.6 STRUCTURAL TRANSFER

Since the natural languages being used are free word ordered, the hybrid approach requires the structural transfer phase. The structural transfer phase uses Naïve Bayes approach to identify the probable grammatical sequence for the target sentence generated by the lexical transfer phase. The features being used in this approach are – sequence of target words and part-of-speech of source word. Before reaching this phase, the syntactic and semantic information are extracted from the text. But based on language analysis, it is found that the syntactic information contributes to the position of the word in the target language whereas, the semantic information does not have any influence on this. Thus, only the syntactic feature is used in the rearrangement phase. The inclusion of syntactic feature in this rearrangement phase has reduced the ambiguity in the grammatical sequence. The maximum probable grammatical sequence is predicted using the alignment table constructed using GIZA++.

5.7 RESULT ANALYSIS

For illustration, consider the source text as - एक सामान्य व्यक्ति का ब्लड प्रेशर 140 - 90 से कम होना चाहिए । (ek saamaany vyakti ka blad preshar 140 - 90 se kam hona chaahie .). The pivot-based approach discussed in previous chapter was applied on this source text and it had slight error in the output text. It is noted that this error can be rectified using the word's sense. Initially, this source text is subjected to word level analysis using morphological analyzer. Since, the input text has only root words and doesn't have any TAM information, there is no change in the input text after morphological analysis. This text is further fed to part-of-speech tagger and the output of the tagger is - एक/QT_QTC सामान्य/JJ व्यक्ति/N_NN का/PSP ब्लड/N_NN प्रेशर/N_NN 140/QT_QTC -/RD_SYM 90/QT_QTC से/PSP कम/QT_QTF होना/V_VAUX चाहिए/V_VAUX ।/RD_PUNC.

The word sense will be disambiguated with the help of latent semantic analysis (LSA). Consider the word "सामान्य (saamaany)" and its identified part-of-speech i.e., adjective. The word and its part-of-speech are used to extract the various possible senses from the Hindi wordnet. The various senses retrieved for the word "सामान्य (saamaany)" from the wordnet are – साधारण (sadharan), नार्मल (normal), सामूहिक (samuhik), सार्वत्रिक (saarvthrik), सादा (saada), आम (aam). All the sentences for these senses are retrieved from the wordnet. Retrieved sentences are mentioned below in the order of identified senses,

S1: यह सामान्य साड़ी है ।

(yah saamaany sari hai .)

S2: शहर की हालत सामान्य हो रही है ।

(shahar kee haalat saamaany ho rahee hai .)

S3: सामूहिक सभा का आयोजन किया गया ।

(saamoohik sabha ka aayojan kiya gaya .)

S4: विज्ञान के नियम सामान्य होते हैं ।

(vigyaan ke niyam saamaany hote hain .)

S5: बाबा आमटे सादा जीवन जीते हैं ।

(baaba aamate saada jeevan jeete hain .)

S6: आम आदमी मँहगाई से परेशान है ।

(aam aadamee manhagaee se pareshaan hai .)

The term-document frequency matrix is constructed using all the above sentences [S1, S2, …, S6] along with the input sentence. Using these 7 sentences, the number of distinct words is identified as 39. Thus, the term-document frequency matrix (A) will be of size (39x7) and is as shown below,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.6}$$

Using singular value decomposition [66], the term-document frequency matrix is decomposed into three different matrices – left singular matrix (L), right singular matrix (R) and singular diagonal matrix (S). The left singular matrix will be of size (39x39), whereas the right singular matrix will be of size (7x7). The singular diagonal matrix will also be of size (7x7), but, it has non-zero values as diagonal elements. These three matrices are as shown below,

$$L = \begin{bmatrix} -0.2244 & -0.1560 & 0.0548 & -0.0176 & \cdot & \cdot & -0.1078 \\ -0.4330 & 0.2353 & 0.0790 & 0.3091 & \vdots & \vdots & 0.0460 \\ -0.2244 & -0.1560 & 0.0548 & -0.0176 & & & -0.0196 \\ -0.2515 & -0.1931 & -0.3467 & 0.0426 & \vdots & \vdots & -0.0071 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \vdots & \vdots & \cdot \\ -0.0560 & 0.0830 & -0.0137 & -0.2584 & & & 0.8237 \end{bmatrix} \tag{5.7}$$

$$R = \begin{bmatrix} -0.8486 & -0.4981 & 0.1326 & -0.04210 & 0.1049 & -0.0260 & 0.0270 \\ -0.2243 & 0.3204 & 0.0078 & 0.0928 & 0.0014 & 0.1387 & -0.9051 \\ -0.2966 & 0.5053 & 0.0347 & 0.4862 & -0.0543 & -0.6103 & 0.2089 \\ -0.1022 & -0.1187 & -0.9721 & 0.1439 & -0.0975 & 0.0167 & -0.0078 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -0.2118 & 0.4238 & -0.0331 & -0.6172 & -0.7006 & -0.0856 & 0.0685 \end{bmatrix} \quad (5.8)$$

$$S = \begin{bmatrix} 3.7815 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.1934 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.4215 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.3890 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.6008 \end{bmatrix} \quad (5.9)$$

The cosine similarity between vectors in right singular matrix and first row of singular diagonal matrix is calculated. The resultant vector after applying cosine similarity is as below,

$$Cosine\ similarity = [0.13 \quad 0.3341 \quad 0.5423 \quad 0.4866 \quad 0.5335 \quad 0.4971 \quad 0.5086] \quad (5.10)$$

The sentence which has the cosine similarity nearer to 0 is closest to the input sentence. The first value in the vector denotes the cosine similarity with the input sentence itself and it is natural that it will be closest to zero. The next smallest value in the vector is 0.3341 which is the cosine similarity value of sentence S1. The word's sense used in S1 is साधारण (saadhaaran) and it is the most matching sense for the word सामान्य (saamaany) according to the context in which it is used.

Table 5.2: Transfer table for Hindi word "saadhaaran"

S. No.	Hindi Word	Part-of-speech Tag	English Word	Frequency of occurrence
1	साधारण	JJ (Adjective)	Simple	4
2	साधारण	JJ (Adjective)	Normal	12

The lexical transfer phase uses the identified semantically similar word साधारण (saadhaaran) instead of the actual word सामान्य (saamaany). The transfer table for the Hindi word "साधारण (saadhaaran)" is as shown in Table 5.2.

For instance, consider the English word (pivot) as “simple” whose frequency of occurrence in the corpus is 36. Out of these 36 occurrences, it was tagged as adjective in 24 instances. Consider the probable preceding word as “A”, which occurred 54 times in the corpus. The word “A” precedes word “simple” in 6 occurrences. Thus, the probable pivot word is found using below expression,

$$\boldsymbol{P}\left(\frac{w_p}{w_s, pos}\right) = \boldsymbol{P}\left(\frac{w_s}{w_p}\right) * \boldsymbol{P}\left(\frac{pos}{w_p}\right) * \boldsymbol{P}\left(\frac{w_p}{w_{p-1}}\right) = \frac{4}{36} * \frac{24}{36} * \frac{6}{54} = \mathbf{0.0082} \qquad (5.11)$$

Considering the pivot word (in English) as “normal” whose frequency of occurrence in the corpus is 24. Word “normal” is tagged as adjective in 12 of its occurrences in the corpus. The word “A” precedes word “normal” in 2 occurrences. The probability of word “normal” as pivot word is calculated below,

$$\boldsymbol{P}\left(\frac{w_p}{w_s, pos}\right) = \boldsymbol{P}\left(\frac{w_s}{w_p}\right) * \boldsymbol{P}\left(\frac{pos}{w_p}\right) * \boldsymbol{P}\left(\frac{w_p}{w_{p-1}}\right) = \frac{12}{24} * \frac{12}{24} * \frac{2}{54} = \mathbf{0.0092} \qquad (5.12)$$

Table 5.3: Transfer table for English word “normal”

S. No.	English Word	Part-of-speech Tag	Tamil Word	Frequency of occurrence
1	Normal	JJ (Adjective)	சாதாரண (Cātāraṇa)	85
2	Normal	JJ (Adjective)	இயல்பாக (Iyalpāka)	10

From the above calculations, the most probable pivot language word for “साधारण/JJ (saadhaaran)” is “normal”. All the other words are also translated to their probable pivot language word and the resultant pivot language sentence is - “a/DT normal/JJ man/N_NN blood/N_NN pressure/N_NN 140/QT_QTC –/RD_SYM 90/QT_QTC less/QT_QTF than/PSP should/V_VAUX be/V_VAUX”. This intermediate text is fed to transfer phase-II and the calculation for translating the word from “normal” to its equivalent Tamil word is as follows,

The word “normal” can be mapped with the word “சாதாரண (Cātāraṇa)” and also with the word “இயல்பாக (Iyalpāka)”. Firstly, consider the probability calculation of word “சாதாரண (Cātāraṇa)” that occurs 116 times in the Tamil corpus. Out of these 116 occurrences, the word “சாதாரண (Cātāraṇa)” occurs as adjective in 100 instances. Assume, the probable preceding word is found as “ஒரு (oru)” and it occurred 23 times in the corpus. The word “ஒரு (oru)” precedes word “சாதாரண (Cātāraṇa)” in 14 instances. The probability of target word “சாதாரண (Cātāraṇa)” is calculated as below,

$$\boldsymbol{P\left(\frac{w_t}{w_p,pos}\right) = P\left(\frac{w_p}{w_t}\right) * P\left(\frac{pos}{w_t}\right) * P\left(\frac{w_t}{w_{t-1}}\right) = \frac{85}{116} * \frac{100}{116} * \frac{14}{23} = 0.384} \tag{5.13}$$

The word “இயல்பாக (Iyalpāka)” occurred 10 times in the corpus and it is tagged as adjective in 5 of its occurrence. In none of its occurrence, it is preceded by the word “ஒரு (oru)”. Thus, the probability calculation for the Tamil word “இயல்பாக (Iyalpāka)” with the given pivot word “normal” will be,

$$\boldsymbol{P\left(\frac{w_t}{w_p,pos}\right) = P\left(\frac{w_p}{w_t}\right) * P\left(\frac{pos}{w_t}\right) * P\left(\frac{w_t}{w_{t-1}}\right) = \frac{10}{10} * \frac{5}{10} * \frac{0}{23} = 0} \tag{5.14}$$

Table 5.4: Analysis of generated target text with its reference text

Generated Text	ஒரு (oru)	சாதாரண (catarana)	நபரின் (naparin)	இரத்த (iratta)	அழுத்தம் (aluttam)	140 – 90	குறைக்க (kuraikka)	வேண்டும். (ventum)
Reference Text	ஒரு (oru)	சாதாரண (catarana)	நபரின் (naparin)	இரத்த (iratta)	அழுத்தம் (aluttam)	140 – 90	குறைக்க (kuraikka)	வேண்டும். (ventum)
Match	1	1	1	1	1	1	1	1

Based on the above calculation in expression (5.8) and (5.9), it is found that the word சாதாரண (Cātāraṇa) will be most probable translation for the word सामान्य (saamaany). The structural transfer phase has identified that the sequence does not need any rearrangements to retain the target language grammar. The final translated text for this input sentence is shown in the Table-5.4 and it is also compared with the reference sentence mentioned in the corpus.

Table 5.5: Comparison of proposed hybrid approach with the proposed word sense-based approach

S. No.	Corpus Size (in number of words)	Word Sense Based Hindi-English-Tamil Machine Translation		Word Sense based Hindi-Tamil Machine Translation	
		Precision (in %)	Recall (in %)	Precision (in %)	Recall (in %)
1	10000	68	58	65	63
2	20000	70	69	76	72.5
3	30000	81	76	87	86.5

From the Table-5.4, the target sentence is found to be same as it has to be with respect to the reference sentence. The proposed hybrid machine translation system has been evaluated using various sentences (restricted to health domain) and the Bilingual Evaluation Understudy (BLEU) score is calculated to be 0.68. Precision and recall gradually increases with respect to corpus size as shown in Table-5.5. The corpus size when increased, further leads to increase in distortion noise and thus producing a poor translation accuracy. It has been found ideally that the unique word should be kept at 30000 and further increase in the corpus size leads to poor translation due to increase in noise.

The proposed hybrid approach has been compared with the word sense-based Hindi to Tamil machine translation system and also with the pivot-based Hindi to Tamil machine translation system. Figures 5.2 and 5.3 show the comparison of these systems in terms of precision and recall. It is very clear from Figure 5.2 that the precision of proposed hybrid machine translation system improves with respect to corpus size, but its precision is lesser when compared with word sense-based Hindi-Tamil MT system which is due to the increase in distortion by the inclusion of pivot language during translation. From Figure-5.3, it is visible that the recall is relatively good compared to the pivot-based Hindi-Tamil MT. But it degrades when compared to the word sense-based Hindi-Tamil MT due to the distortion caused by the introduction of pivot language in between the source and target language translation.

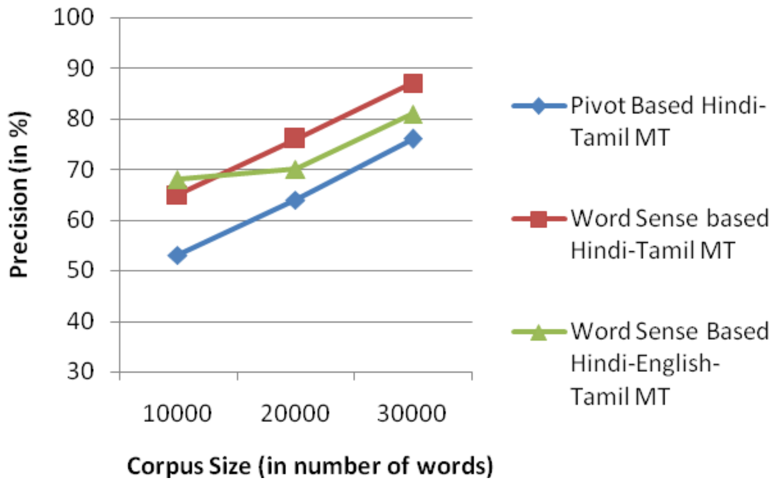

Figure 5.2: Comparison of Hindi to Tamil Machine Translation in terms of precision

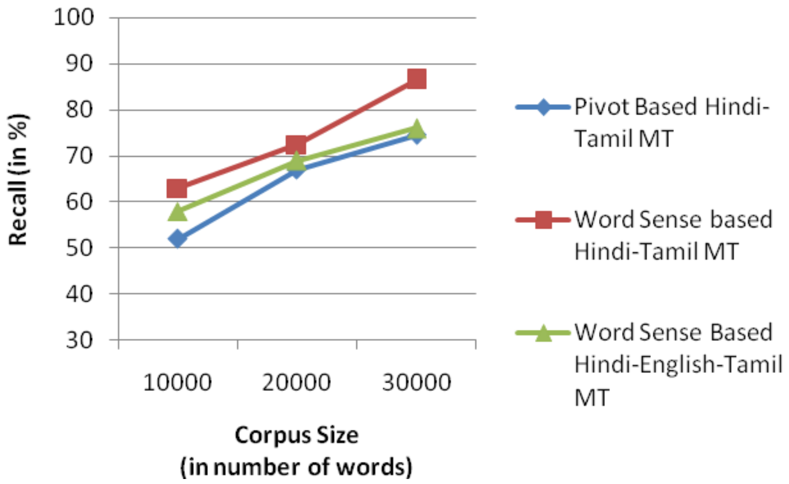

Figure 5.3: Comparison of Hindi to Tamil Machine Translation in terms of recall

The proposed system is also compared with the other statistical machine translation system in terms of BLEU (Bilingual Evaluation Understudy) score, which is shown in Table-5.6. The hybrid approach is found to have a BLEU score of 0.7637 and it is also noted that the BLEU score has improved by few percentage when compared with the pivot-based approach discussed in chapter-4.

Table 5.6: Comparison of various statistical machine translation system using BLEU score

S. No.	**Methodology**	**Source language**	**Target language**	**BLEU Score**
1	Statistical machine translation [67]	English	Bahasa Indonesia	0.2287
2	Lemma translation [68]	Japanese	Indonesian	0.1282
3	Lemma translation [68]	Indonesian	Japanese	0.1723
4	Proposed pivot-based approach	Hindi	Tamil	0.7394
5	Proposed hybrid approach	Hindi	Tamil	0.7637

5.8 CONCLUDING REMARKS

It is evident from the results of the hybrid system that the noise introduced in a pivot-based machine translation system has been reduced by the use of word sense disambiguation along with the pivot-based system. But as compared with the word sense-based Hindi to Tamil machine translation system, the performance of hybrid system is still poor. This is due to the deviation that is happening by the use of pivot language and the language specific properties. Hindi and Tamil language are morphologically rich language when compared with English language. Thus, the morphological structure of words differs between source language, target language and pivot language. This difference leads to loss of semantics during translation and therefore, produces a poor translation than the one generated by direct statistical machine translation (without pivot). These bottlenecks can be addressed by the use of deep learning architecture to capture the semantics and syntactic features of the languages to perform machine translation. The deep learning approach is the topic of the next chapter.

CHAPTER 6

A DEEP LEARNING APPROACH FOR HINDI TO TAMIL MACHINE TRANSLATION

6.1 INTRODUCTION

The hybrid approach on Hindi to Tamil machine translation discussed in last chapter makes use of pivot language and it performs translation using word's semantic feature. But, the accuracy of machine translation was not that good due to the involvement of three natural languages. This also leads to semantic distortion in the translated text. The poor improvement in the accuracy of hybrid machine translation was due to various mappings such as many to one mapping, many to many mapping and in an extreme case of one to one mapping. In Hindi, multiple words are used along with a main word and these additional words contribute to the tense, aspect and modality (TAM) information. For example, the phrase "जा रहा है (ja raha hai)" has the main verb "ja (means – go)" and it also has "raha hai" that contains the present tense and masculine information in it. While translating this phrase to Tamil, the equivalent Tamil word is simply "போகிறான் (pōkiṟāṉ)". Thus, these words along with its root word must be mapped with the single word of the target language. Because of this, there is many-to-one mapping between Hindi and Tamil languages. The Hindi phrase "कैसे है (kaise hai)" is mapped with the Tamil phrase "எப்படி இருக்கிறாய் (eppaṭi irukkiṟāy)" and this is an example of many-to-many mapping. In case of one-to-one mapping, there are direct mapping between two words based on the context in which it is used. Example word for this one-to-one mapping is "पूजा (puja)". In all these cases, if this information is not conveyed to the target language through the intermediate pivot language, then the loss of semantics occurs. There are cases where several words in Hindi are mapped to a single word in English such as, the words - दादा (dada), नाना (nana) are mapped to the word "grandfather" and दादी (dadi), नानी (nani) are mapped to "grandmother". If this English translated word is being used for translation to Tamil, then there is definitely loss in semantic information. Thus, the semantic distortion is a major issue in hybrid approach. In order to handle this issue, a deep learning approach for Hindi to Tamil machine translation is being proposed which is shown in Figure 6.1. There are various deep learning networks such as convolution neural network, recurrent neural network etc. A convolution neural network extracts out the required features from the input data by multiplying it with a filter vector. The features extracted will be in different dimensions. To improve the scalability of this network, these features are further mapped based on their region of existence in the input. Since natural language text are sequential in nature, the recurrent neural network is preferred over a convolution neural network. Recurrent neural network basically uses a feedback mechanism to handle sequential data. The proposed deep learning approach learn features of both languages by considering the syntactic and semantic information in the parallel text used for training. The syntactic feature helps in establishing grammatically correct translation. The semantic information will be helpful in predicting the target text based on the contextual information. By keeping these features into consideration, the approach is designed with the following phases,

i. Word Embedding
ii. Encoder
iii. Attention network
iv. Decoder

Both syntactic and semantic features are embedded into a vector with the help of word embedding. The word embedding is performed using continuous bag-of-words (CBOW) model [69][70]. The continuous bag-of-words model basically maps the word with its neighboring words to generate a vector for it. These neighboring words helps in extracting the syntactic and semantic feature from the sentence. The vectors of various sentences in the parallel bilingual corpus are generated and are fed to the deep learning approach for training purpose. The deep learning approach basically maps the embedded vector in source language with its corresponding embedded vector in target language. But, there is difference in the size of vector in source and target language due to the difference in length of parallel sentences being used. The sequence to sequence approach will be more suitable to handle this difference. In order to learn the features between the vectors, a recurrent neural network is being used in this approach.

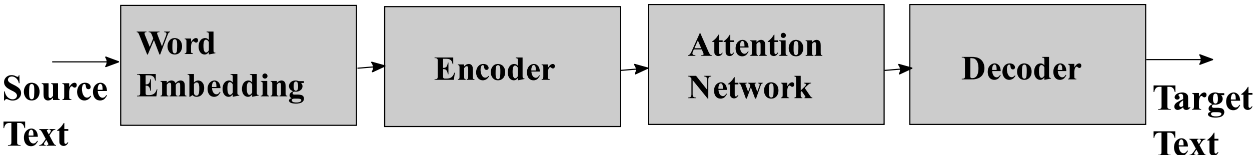

Figure 6.1: Proposed deep learning approach for MT

This approach makes use of an encoder and a decoder to perform machine translation. The encoder encodes the information from source vector to an intermediate fixed length vector. This fixed length intermediate vector is fed to a decoder which generates the target vector. The encoder and decoder are designed using a recurrent neural network (RNN) [71].

In natural languages, there is need for contextual information of word from one sentence to another sentence that occurs later in the corpus. Thus, there is need for keeping a copy of the vector information until it is required for any other sequence. For this purpose, long-short term memory neural network (LSTM), a special type of recurrent neural network, was introduced in the encoder and decoder modules. The input and output vector used for training a sequence to sequence model is generated using a trigram model. The trigram model is found to learn syntactic and semantic features from the sentences. Since the vectors are generated using these features, the accuracy of translation is influenced by these features too.

To maintain the grammatical correctness in the generated target text, an attention network is used in between encoder and decoder. The attention network learns to map the word's positional information by using the encoded vector as input and decoded vector as output. This network is used to detect the influence of encoded vector on the target vector.

6.2 PROPOSED DEEP LEARNING APPROACH

6.2.1 WORD EMBEDDING

Word embedding is a feature learning approach that maps the words or phrases on a continuous vector space. The words that are related based on its sense are placed at the neighboring points in this vector space. The proposed deep learning approach requires both the syntactic and semantic features of the two languages. This leads to extraction of these features from the source text and target text. An appropriate approach to perform this mapping is – continuous bag-of-words (CBOW) model [69][70]. The CBOW model makes use of a neural network which learns the features from the training corpus and generates the vectors based on the neighboring word information. If the number of neighboring words being used is ‘m’, then the continuous bag-of-words model predicts the word at nth position (W_n) using all the ‘m’ preceding words. In deep learning approach, a trigram model is being used to capture the syntactic and semantic features. A trigram model is found to have more appropriate mapping as compared with the n-gram model. In a trigram model, the CBOW approach predicts the word at position 3 using the information about the words at position 1 and 2. The model uses a neural network to capture the features from the words and it is shown in Figure-6.2. It has an input layer, hidden layer and output layer. The size of input layer and output layer is equal to the size of vocabulary. The size of hidden layer is kept at an arbitrary value which represents the dimension required to embed the words.

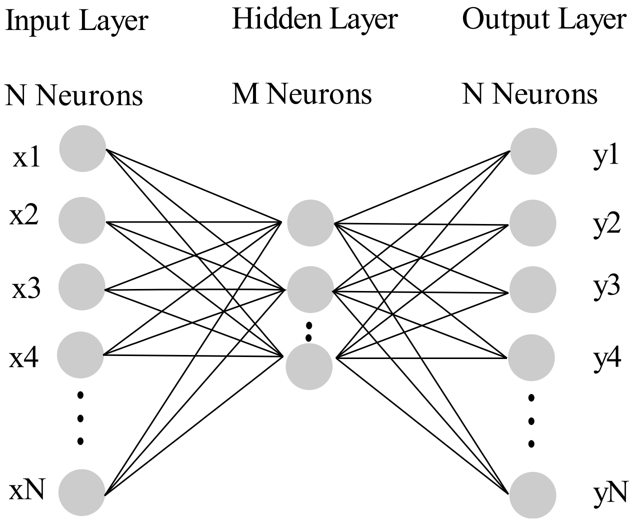

Figure 6.2: CBOW model for word embedding

A one hot encoding is used to represent the variables in binary vector form. For each of the words along with its neighboring words, a one hot vector is generated using this encoding. The number of neighboring words to be considered is decided based on the context window. For example, if the size of context window is one, then every word along with its preceding word is considered during encoding. The generated one hot vector is fed as input to continuous bag-of-words model. Output layer obtains a list of probabilities generated based on the one hot vector. Output layer has softmax layer which is used to find the sum of these probabilities. The output from this network embeds the word vector based on the context in which it is being used. There is a weight matrix (W_{IH}) in between input and hidden layer in this model. The input one hot vector is multiplied by this weight matrix (W_{IH}). This vector is further multiplied by the weight matrix (W_{HO}) that occur in between hidden and output layer. Cross entropy is used to detect the distance between the embedded vectors. According to the distance measure, the word vectors are updated using gradient descent approach [72]. The gradient descent approach tries to minimize the distance measured by varying the weights of the network according to the learning rate.

For example, consider the input sentence as - चुइंग गम चबाने से लार बनती है। (chyuing gam chabaane se laar banatee hai.). Assuming the context window size as 1, the one hot vector for every word in this sentence is as shown in Table 6.1.

Table 6.1: One hot encoding for words in Hindi text – "chyuing gam chabaane se laar banatee hai."

S.No.	**Input Word**	**Output Word**	**चुइंग** (Chyuing)	**गम** (Gam)	**चबाने** (Chabaane)	**से** (Se)	**लार** (Laar)	**बनती** (Banatee)	**है।** (hai.)
1	चुइंग	गम	1	0	0	0	0	0	0
2	गम	चुइंग	0	1	0	0	0	0	0
3	गम	चबाने	0	1	0	0	0	0	0
4	चबाने	गम	0	0	1	0	0	0	0
5	चबाने	से	0	0	1	0	0	0	0
6	से	चबाने	0	0	0	1	0	0	0

7	से	लार	0	0	0	1	0	0	0
8	लार	से	0	0	0	0	1	0	0
9	लार	बनती	0	0	0	0	1	0	0
10	बनती	लार	0	0	0	0	0	1	0
11	बनती	है।	0	0	0	0	0	1	0
12	है।	बनती	0	0	0	0	0	0	1

Assuming size of hidden layer as 3, the dimension of weight matrix (W_{IH}) will be (7x3) and the dimension of (W_{HO}) will be (3x7). Consider, the randomly generated weight matrices as,

$$W_{IH} = \begin{bmatrix} 0.1 & 0.8 & 0.6 \\ 0.2 & 0.9 & 0.7 \\ 0.3 & 0.1 & 0.8 \\ 0.4 & 0.2 & 0.9 \\ 0.5 & 0.3 & 0.1 \\ 0.6 & 0.4 & 0.2 \\ 0.7 & 0.5 & 0.3 \end{bmatrix}, W_{HO} = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 \\ 0.8 & 0.9 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 \\ 0.6 & 0.7 & 0.8 & 0.9 & 0.1 & 0.2 & 0.3 \end{bmatrix}$$

Consider the input word as “चुइंग (Chyuing)” and its one-hot encoding is [1, 0, 0, 0, 0, 0, 0]. Input to the hidden layer IN_H is calculated by multiplying the transpose of input vector with the weight matrix W_{IH}.

$$IN_H = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] * \begin{bmatrix} 0.1 & 0.8 & 0.6 \\ 0.2 & 0.9 & 0.7 \\ 0.3 & 0.1 & 0.8 \\ 0.4 & 0.2 & 0.9 \\ 0.5 & 0.3 & 0.1 \\ 0.6 & 0.4 & 0.2 \\ 0.7 & 0.5 & 0.3 \end{bmatrix} \tag{6.1}$$

$$IN_H = [0.1 \quad 0.8 \quad 0.6] \tag{6.2}$$

The matrix (IN_H) is fed to the hidden layer and it is multiplied with weight matrix W_{HO} to generate the output of hidden layer (O_H)

$$O_H = [0.1 \quad 0.8 \quad 0.6] * \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 \\ 0.8 & 0.9 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 \\ 0.6 & 0.7 & 0.8 & 0.9 & 0.1 & 0.2 & 0.3 \end{bmatrix} \tag{6.3}$$

$$u = \ O_H = [1.01 \quad 1.16 \quad 0.59 \quad 0.74 \quad 0.35 \quad 0.5 \quad 0.65] \tag{6.4}$$

Predicted output vector of jth word (O_j) is calculated by applying softmax function on the hidden layer output (u). Softmax is used to predict the posterior distribution of words. It is expressed as,

$$O_j = \frac{e^{u_j}}{\sum_{i=1}^{n} e^{u_i}} \tag{6.5}$$

Where,

u_k – k^{th} value of the vector 'u'

O_j − output of j^{th} word in the vocabulary

n − vocabulary size

Each of the word's probability is predicted using the softmax function mentioned in expression (6.5). These probabilities are used to populate the predicted output vector (O_{pred}). The predicted output vector the input sentence will be as given below,

$$O_{pred} = [0.1853 \quad 0.2153 \quad 0.1217 \quad 0.1414 \quad 0.0958 \quad 0.1112 \quad 0.1293] \tag{6.6}$$

The loss between the predicted output $(O_{pred})$and the desired output (O_{act}) is calculated. Since the word "गाम (Gam)" occurs after the input word "चुइंग (Chyuing)", the desired output is considered as the one-hot encoding of the word "गाम (Gam)", i.e., [0, 1, 0, 0, 0, 0, 0]. To reduce loss, backpropagation algorithm is used to update the weights on the network.

6.2.2 SEQUENCE TO SEQUENCE MODEL

In a sequence to sequence approach, the deep learning network learns the sequence at (n+1)th positions provided the ‘n’ sequences are fed to it. In case of machine translation, the sequence to sequence approach predicts the target sentence based on the sequence of words in source text. Sequence to sequence machine translation approach has a combination of encoder and a decoder, as shown in Figure-6.3. In this figure, the length of source language sentence is kept as ‘N’ and the target language sentence length is kept as “M”. The source language words are denoted as S_1, S_2, S_3, …, S_N. Words in target language are denoted as T_1, T_2, T_3, …, T_M. The word S_N encodes its information along with the information passed to it from (N-1)th encoder unit. The encoder encodes the features from input vector into an intermediate fixed length vector and this fixed length vector is used by the decoder to generate the output vector. The target word T_M is predicted by using the decoder output T_{M-1}. The training of encoder and decoder should be performed in parallel manner using the vector of source sentence and its equivalent target sentence vector. The source and target language vectors are generated using continuous bag-of-words (CBOW) model on the respective monolingual corpus, so that, the syntactic and semantic information stored in the monolingual corpus can be captured to encode into the vectors. These vectors are further used in the sequence to sequence model to learn the features between these vectors.

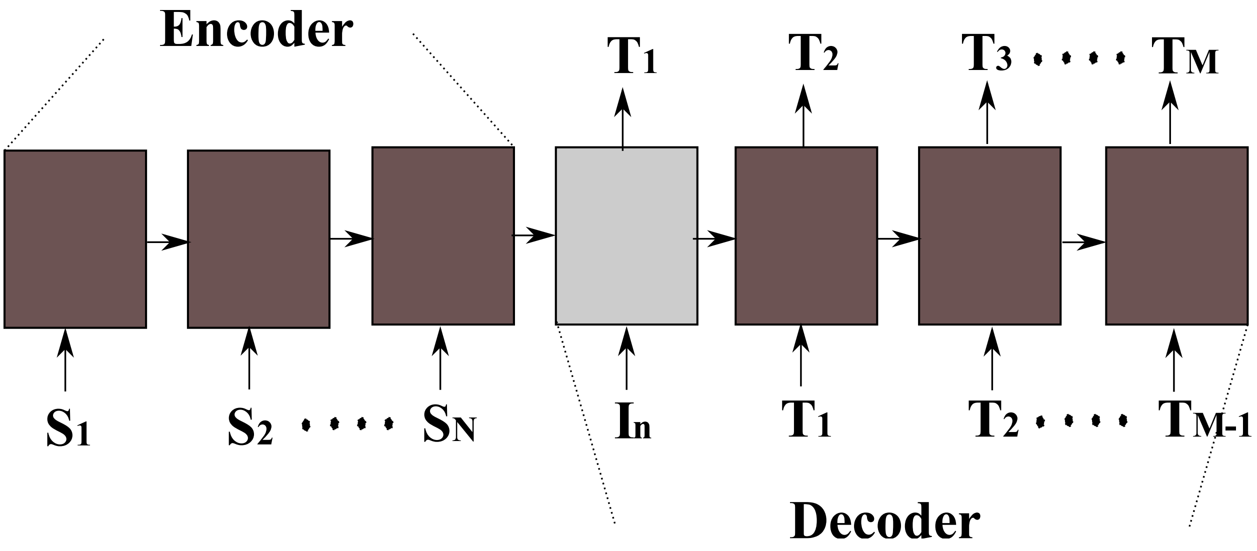

Figure 6.3: Sequence to Sequence model for machine translation

This model learns features between vector using a recurrent neural network (RNN). Recurrent neural network stores the output of previous computation at time *(t-1)* and it is fed to the current vector processing at time '*t*'. But, in natural languages, the context of a word has dependency on any other sentence that has the same word in it. The LSTM network will be more suitable for natural language applications such as machine translation. Hence, sequence to sequence model uses long-short term memory (LSTM) network. Long-short term memory network can learn the long-term dependencies between the sentences. This network keeps a cell memory to store information for a certain period of time. The information stored retains contextual information and syntactic information. This stored information is further used while processing a vector which has dependency on it.

6.2.3 LONG-SHORT TERM MEMORY (LSTM) NETWORK

A recurrent neural network (RNN) is used to learn the sequences by storing the computation of previous output vector sequence and uses this stored vector during the computation of current sequence [71]. Due to this, it learns the relationship between two consecutive sequences. The sequence mapping plays a major role in case of natural language processing. Thus, recurrent neural network is most appropriate for most of the natural language applications. In certain applications, there is need for finding the relation between the word in n^{th} sentence with the same word in $(n-i)^{th}$ sentence. This emphasizes on learning the relationship between sequences which are not in consecutive manner. The LSTM network uses LSTM cells to store the features learnt from the sequences. It makes use of gates to predict the usage of stored vector for the computation of current sequence. The encoder and decoder in this sequence to sequence model is developed using LSTM network and is shown in Figure-6.4. The encoder learns the relation between words and encodes them into an intermediate vector. This intermediate vector contains the syntactic and semantic information of words used in the source sentence. The decoder uses this intermediate vector to decode and generate the sentence on the target language.

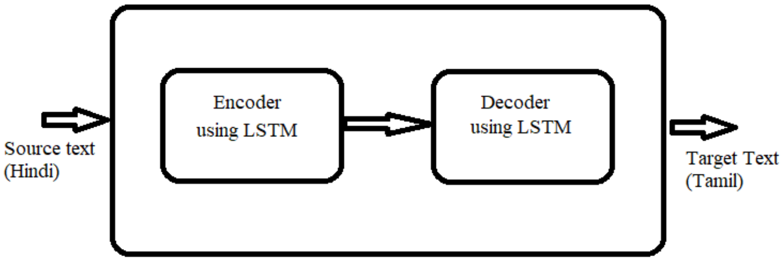

Figure 6.4: Sequence to Sequence model

6.2.4 ENCODER AND DECODER

The main aim of the encoder module is to generate an intermediate vector which has encoded information about the input sentence. The LSTM network is used to handle long-term dependencies between vectors. The network has a LSTMCell to store the vector information. There are four different gates used to protect and control this LSTMCell. The four different gates are – input gate, out gate, cell gate and forget gate. The gates act like an activation function in a neural network. The input gate and cell gate are used to decide which value will be stored in the memory. The value that is stored in LSTMCell is retained or not is decided using the forget gate. The out gate decides about whether the value is required or not for the current sequence processing. The value at LSTMCell has the information about which value in the output vector has influence on the current sequence. The influence is measured in between -1 to 1. The tanh activation function maps the value in this range and thus helps to extract the influence of vector values on the current sequence. At time *t*, the output of hidden state h_t is calculated using the expression given below,

$$h_t = o_t \tanh c_t \tag{6.7}$$

Where, o_t- out gate and c_t- value at LSTMCell

The LSTMCell value (c_{t-1}) at time *(t-1)*, new cell value g_t, forget gate value f_t and input gate value i_t are used to find the value at LSTMCell c_t. The below mentioned equation is used to find the value to be stored at LSTMCell,

$$c_t = f_t * c_{t-1} + i_t * g_t \tag{6.8}$$

Where, f_t – forget gate and i_t- input

The forget gate value is multiplied, in element-by-element manner, with the previous LSTMCell value. This is performed to predict the need of previous cell value in the current cell value prediction. The input gate value and new cell value are multiplied to find the need for storing it in the LSTMCell value c_t.

The value stored in the hidden layer at *(t-1)* and the current input vector (x_t) are used to find the out-gate value (o_t). Depending on the current input vector, the out gate will be dependent on either the current input vector or on the previous hidden layer value h_{t-1}. Thus, there is assignment of weights to x_t and h_{t-1} during the calculation of out-gate value (o_t). The weight that decides about the input vector is calculated based on weights assigned in between the input gate and out-gate. The weights that are assigned in between hidden layer and out-gate are used for deciding the influence of previous hidden layer value. The values of gates should be in the range of 0 to 1 and thus, a sigmoid activation function is applied over this vector calculation. It is mathematically expressed as,

$$o_t = sigmoid(w_i^o * x_t + b_i^o + w_h^o * h_{t-1} + b_h^o) \tag{6.9}$$

Where,

w_i^o- Weight vector between input layer and out gate

w_h^o- Weight vector between hidden layer and out gate

b_h^o- Bias between hidden layer and out gate

b_i^o- Bias between input layer and out gate

The out-gate value and LSTMCell value at time *(t-1)* are major backbone of the LSTM network and this is used to calculate the current LSTMCell value. But the out-gate value and LSTMCell value are dependent on the other gate values such as input gate, forget gate and new cell value. Each of these gate values are calculated using the following expression,

$$g_t = \tanh(w_i^g * x_t + b_i^g + w_h^g * h_{t-1} + b_h^g) \tag{6.10}$$

$$f_t = sigmoid\left(w_i^f * x_t + b_i^f + w_h^f * h_{t-1} + b_h^f\right) \quad (6.11)$$

$$i_t = sigmoid(w_i^i * x_t + b_i^i + w_h^i * h_{t-1} + b_h^i) \quad (6.12)$$

For illustration, consider the Hindi sentences mentioned below,

Hindi Text: बेमौसम की सब्जी न खाएँ । मिटटी के बरतन में रखा पानी पिएँ ।

(Bemausam ke sabjee na khaen. Mittee ke barten mein rakha paanee pien.)

English Equivalent: Do not eat non-seasonal vegetables. Drink water kept in earthen pots.

In the above Hindi texts, the second sentence has no dependency with the information from first sentence. This is due to difference in the words being used in these sentences. During translation of the second sentence, there is no need to keep the contextual information from the first one. Thus, the forget gate will enabled to discard the vector information about the first sentence. The value of forget gate should be in the range from 0 to 1. To forget a particular information, the vector value of forget gate is made as 0 or else as 1. Sigmoid function will be useful in mapping the values in the range of 0 to 1.

Consider the next set of Hindi texts for the discussion about input gate and output gate,

Hindi Text: बेमौसम की सब्जी न खाएँ । समय से खाना खाएँ ।

(Bemausam ke sabjee na khaen. samay se khaana khaen.)

English Equivalent: Do not eat non-seasonal vegetables. Eat food in time.

The above two sentences share contextual information between them. The vector information of first sentence will be helpful during the translation of the second one. Thus, it has to be stored in the network for later use. The input gate will be helpful in achieving this. Its functionality is same as the forget gate but, it acts as a filter to extract only those vectors which will be useful later. If some vector has useful information for the current translation process, then the sigmoid function maps it to 1 or else maps it to 0. The filtered vector information is multiplied by the current cell value to identify the relation between the current sentence and its previous. This information is added with vector information that is stored in the cell state.

In the above two Hindi sentences, all the information stored in first sentence are not useful while translating the second one. The out gate will be helpful in filtering useful information from the stored ones. In the first sentence, vector values of all the words except "खाएँ (khaen)" are not useful while translating the second sentence. Thus, the out gate suppresses those values which are not useful for the current processing.

The decoder module makes use of a LSTM network for learning the features that maps the intermediate vector with the output vector. The intermediate vector which is fed to decoder is generated from the encoder. During training, the intermediate vector and target vector are used to learn the features. Once the training is completed, the decoder module is able to generate the target vector based on the intermediate vector fed to it. The LSTMCell that stores the vector information in a decoder is decided using the intermediate vector and its computed gate values.

6.2.5 ATTENTION MECHANISM

The decoder module of a sequence to sequence model described above makes use of the intermediate vector generated by the encoder. The main purpose of encoder is to encode the feature information into the intermediate vector. So that it can be used by the decoder later on. If the input sequence is of length 'n', then the encoder encodes the information at 1st position and passes the encoded information to the next sequence for processing. Likewise, every sequence is processed with the help of the encoded information from its previous position. Thus, the intermediate vector stores the contextual information of all its previous sequences. It helps the decoder to generate the sequence of words in target language using the contextual information. In case of Indian language pairs such as Hindi and Tamil, the contextual information of target language has dependency with any of the word mentioned in the source text. This dependency is not fixed for sequence of words due to the free word order nature of both these languages. This leads to inefficiency in the performance of sequence to sequence network as the contextual information stored in intermediate vector is not sufficient to perform the decoding process. Also, the target text being generated in the sequence to sequence model will not be grammatically correct and thus, there is need for special rearrangement phase. These issues can be handled with the help of attention mechanism [73]. The attention mechanism is a method to identify the influence of the encoded sequence with respect to the target sequence. This influence can be learnt with the help of a neural network in between the encoder and decoder module. Based on the influence of encoded sequence on target sequence, a weighted vector is generated by multiplying the encoded output with its hidden layer weights. The hidden layer weights are calculated using a feedforward network which uses the encoded sequence as input and decoder's output as its output.

6.3 RESULTS AND DISCUSSION

The proposed sequence to sequence model was developed using pytorch with tensorflow at the backend. Tensorflow is an open source platform used for developing deep neural networks. The proposed model learns the features from the input vector and target vector. These features are used to generate the target text based on the input vector fed to it. To make the model learn the features in an efficient way, there is need for huge amount of corpus. But, the languages Hindi and Tamil has low resources when compared with the resources of English language. To handle this issue, the language specific features were explored, and the free word order feature will be helpful in achieving a good result. Hindi language is partially free word ordered and Tamil language is fully free word order language. Due to this feature, a sentence in Tamil and Hindi can be shuffled in different combinations to generate variants of the given sentence. Since Hindi language is partially free word order language, all the combination generated will not be grammatically correct. Thus, there is need for verifying grammatical correctness of the text being generated. Parsing the sentences will be helpful in checking the grammatical correctness of it. In this proposed approach the Hindi shallow parser [74] is used to verify the correctness of generated Hindi sentence. The Hindi parser verifies the grammar by parsing the tagged text fed to it. In this way, the valid variants of Hindi text are generated along with its Tamil sentences and are maintained in the training dataset. Similarly, the Tamil sentences can also be shuffled but there is no necessity for verification of grammar in it. This is due to the fully free word order nature of the language.

This complete dataset is fed to the sequence to sequence model for training purpose. The model is trained using the vector generated from source language and the vector from target language. The features of parallel sentence are extracted during the training process. The training process is continued until the error between the target vector and generated vector is as low as possible. During training, there was overfitting issue which leads to poor performance by the overall network. Due to overfitting, the target text being generated was found to have more fluctuations and it signifies that the network was not able to learn the features properly. To train the network without overfitting issue a dropout regularization mechanism [37] was introduced. The dropout mechanism nullifies a random percentage of neurons and tries to learn feature from other neurons. The optimum percentage of dropout has to be chosen. If not, then there will be case of underfitting in the network. The proposed model was analyzed with various dropout percentage and an optimal percentage value is found to be in the range of 20% to 60%. Since there is an encoder module and a decoder module, there is need for analyzing the dropout percentage in both these modules such that the performance of the overall system is good. The ideal dropout percentage for encoder is found be 20% and the ideal dropout percentage for decoder is 60%. The following are parameters that was used for sequence to sequence model,

Number of epochs = 22
Learning rate = 0.01
Hidden layer size = 2
Dropout = 0.2 (in encoder) and 0.6 (in decoder)

After 22 epochs, the feature learning by model gets saturated and the necessity for training becomes negligible. The learning rate is kept at 0.01 and the accuracy of the model fluctuates when the learning rate is made a 0.1. This is due to rapid change on the weights and thus the model was not able to learn the features. The number of hidden layers used in the proposed approach is two. The performance of proposed model is comparatively better as compared with the model having more than two hidden layers.

The proposed sequence to sequence model has been tested on various input sentences and a few sample generated target sentences are shown in Table-6.2 and Table-6.3 below. It also has the comparison of the generated Tamil text with the reference text in the corpus.

Hindi Text: चुइंग गम चबाने से लार बनती है।

Transliterated text: chyuing gam chabaane se laar banatee hai.

English equivalent: Saliva is formed by chewing the chewing gum.

Table 6.2: Comparison of generated Tamil text for input text “chyuing gam chabaane se laar banatee hai.”

Target Text (generated):	சூயிங்கம் (cuyinkam)	மெல்லுவதினால் (melluvatinal)	உமிழ்நீர் (umilnir)	உற்பத்தியாகின்றது. (urpattiyakinratu)	<EOS>
Reference Text	சூயிங்கம் (cuyinkam)	மெல்லுவதினால் (melluvatinal)	உமிழ்நீர் (umilnir)	உற்பத்தியாகின்றது. (urpattiyakinratu)	<EOS>
Matching	1	1	1	1	1

Hindi Text: प्याज के सेवन से आँखों की ज्योति बढ़ती है ।

Transliterated text: pyaaj ke sevan se aankhon kee jyoti badhatee hai .

English equivalent: Eye sight is improved by consuming onion.

Table 6.3: Comparison of generated Tamil text for input text “pyaaj ke sevan se aankhon kee jyoti badhatee hai .”

Target Text (generated):	கண்களில் (kankalil)	சாப்பிடுவதினால் (cappituvatinal)	கண்களில் (kankalil)	ஒளி (oli)	அதிகரிக்கிறது . (atikarikkiratu)	<EOS>
Reference Text	வெங்காயம் (venkayam)	சாப்பிடுவதினால் (cappituvatinal)	கண்களில் (kankalil)	ஒளி (oli)	அதிகரிக்கிறது . (atikarikkiratu)	<EOS>
Matching	0	1	1	1	1	1

The proposed sequence to sequence model was tested with various test set. The generated target sentences were evaluated using Bilingual evaluation understudy (BLEU) score. The Table-6.4 shows the analysis of BLEU score with different training and testing pairs. The BLEU score calculated for this proposed system is the average of sentence level BLEU score. The BLEU score is found to improve with increase in the percentage of training set being used. The accuracy of proposed model is good when the training and testing corpus ratio is kept at 80:20.

Table 6.4: Result analysis of Neural Machine translation

S. No.	Training/Testing Corpus Size (in %)	BLEU Score
1	60/40	0.7037
2	70/30	0.7234
3	80/20	0.7588
4	90/10	0.6628

The neural machine translation system is also evaluated at different runs by keeping the ratio of training and testing pair as 80:20 as shown in Table 6.5. At each of the runs the training and testing set has been changed and are chosen at random. It is also found that the neural machine translation system has accuracy better than the statistical machine translation system.

Table 6.5: Result analysis of Neural Machine translation on different runs

Number of runs	BLEU score
1	0.7478
2	0.7176
3	0.6784
4	0.7118
5	0.7588
6	0.7124
7	0.7022
8	0.6914
9	0.7156
10	0.7211

6.4 CONCLUDING REMARKS

The word embedding is performed using a continuous bag-of-words model and it is found to capture the semantics in the words. This in turn helped in improving the accuracy of the translation using the sequence to sequence model. Since Hindi and Tamil language are morphologically rich, there is need for semantic mapping which is made using this approach. The results are found to be far better than any state-of-art method for these two languages. It is found that BLEU score is 0.7588 and it can be improved further by using a properly aligned parallel corpora.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 CONCLUSIONS

In today’s multicultural business world, there is involvement of more natural languages for establishing communication in the business environment. Due to globalization, there is need for machine translation system to assist in communication between two different organization. Thus, the demand for translation system on various language pairs has increased. It can also help in improving the communication between people from different origin. The development in machine translation also promotes the re-establishment of various business activities according to the national resources available for the business. For example, a Canadian based MNC can establish its office in India if they get proper translation services. To achieve this, there is prime need for developing machine translation system with more accuracy. In the current scenario, the machine translation system between Indian languages are still in need of development due to its poor accuracy. Similarly, there is need for machine translation services between prominent languages of northern and southern India for more business/technical interactions.

Developing a machine translation system between Indian language pair is the primary and major issue that needs to be addressed. The Indian languages such as Hindi and Tamil have poor resource availability. Existing translation system on this language pair has focus towards syntactic features of the languages. But, there is need to consider the semantic features of the languages too. Thus, a combination of both syntactic and semantic feature will provide a more accurate machine translation.

The next major challenge was to generate the parallel corpus that is required for a machine translation system. Since, the availability of resources in Hindi and Tamil is very poor, there was need for some intermediate pivot language to assist in the translation. This research work reports about one such pivot-based approach which uses English as the pivot language due to its vast resource availability.

One more major challenge in the research was to improve the overall accuracy by making use of the language specific features. Hindi is morphologically rich and partially free word ordered language. Whereas, Tamil is morphologically rich but fully free word ordered language. Since both the languages are morphologically rich, the tense, aspect and modality information are stored along with the root word. During translation, this information also plays major role. Apart from this, the word order also contributes to an accurate translation. All these language specific features have to be extracted and used for improving the performance of a translation engine.

The target sentence being generated should follow the grammar rules of target language. Grammar followed by Hindi and Tamil languages are in the subject-object-verb (SOV) form. The Hindi sentences are partially free word order whereas, Tamil sentences are fully free word order. This word order feature increases the challenge in mapping the target text according to its grammar rules.

The following are list of observations and contributions in this research on Hindi to Tamil machine translation,

1. Word sense-based approach for Hindi to Tamil machine translation was proposed, which considers both the syntactic and semantic feature of both languages. Syntactic features contribute to the mapping of sentences with its probable target text. But the semantic feature provides more detailed information about the most appropriate target word for a given source word. It is observed that the approach has issue with the syntactic information being retrieved. This is due to poor performance by the HMM based part-of-speech tagger. Even the amount of resources needs to be increased for better performance of this approach. But, both Hindi and Tamil are poor resource languages.
2. A pivot-based approach for translation was proposed to handle the low resource issue. Since, English has vast resources as compared with Hindi and Tamil, English is used as a pivot language. To improve the performance of syntactic process, a multilayer perceptron based neural network was used to perform part-of-speech tagging. The multilayer perceptron based neural network extracts the syntactic information from the

input sentence and this information is used during the translation process. But Hindi and Tamil are morphologically rich languages as compared with the morphology of English. Due to which there is loss of semantics and it is being handled by the introduction of sense disambiguation phase in a pivot-based approach.

3. A hybrid approach was proposed to handle the semantic distortion that occurred in a pivot-based approach. This approach uses word sense disambiguation and the identified senses aid in the translation process. The syntactic information extracted using multilayer perceptron tagger is also used in this hybrid approach. The improvement in performance of Hindi to Tamil translation by the use of hybrid system saturates after particular threshold on corpus size. Due to difference in word morphology in Hindi, English and Tamil, there is introduction of semantic distortion during translation. Thus, the actual information gets lost during translation.
4. In order to improve the translation accuracy further, a deep learning approach was proposed. This approach learns the features from the parallel text fed to it. The parallel text was embedded into a vector without losing the syntactic and semantic features of the languages. Deep learning approach uses attention mechanism to extract the features that can help in sentence rearrangement process.

The machine translation approaches were evaluated using Bilingual evaluation understudy (BLEU). The BLEU score of proposed approaches have been calculated and are listed in the Table 7.1.

From the table, it is visible that the performance of all the approaches except pivot-based approach is better. This is due to semantic distortion that occurred by the usage of three different languages. The performance of hybrid approach is better when compared with all other approaches. But, its performance does not improve after a particular threshold value even with increase in the corpus size. Out of all the approaches, the deep learning approach has performed better when compared with word sense-based approach and pivot-based approach, and it can be improved further with increase in corpus size.

Table 7.1: Comparison of various proposed machine translation approaches in terms of BLEU score

S. No.	Approach	BLEU score
1	Word sense-based approach	0.6800
2	Pivot-based approach	0.5400
3	Hybrid approach	0.7637
4	Deep learning approach	0.7588

7.2 FUTURE WORK

The significant and prominent improvements that can further increase the accuracy of Hindi to Tamil machine translation system are listed below:

1. The word sense-based Hindi to Tamil statistical machine translation system can further be improved by the introduction of more accurate part-of-speech tagger and word sense disambiguation. The increase in corpus for this approach may also improve the overall accuracy of the system.
2. A multilayer perceptron-based part-of-speech tagger was designed using the statistical features. The performance of tagger with one hidden layer is found to be better as compared with the network having more than one hidden layer. The statistical feature being used for training the network is based on a bigram language model.
3. Pivot based statistical machine translation can also be improved by the introduction of more accurate part-of-speech tagger. The performance of this approach may be improved by using a pivot language which has high resource availability and has some relation with either the source or the target language. Thus, the use of a different pivot language may be explored.
4. The hybrid approach uses the word sense module along with the pivot-based approach. This approach can also be improved by using a pivot language which has relationship with either the source or the target language. Even the improvement in accuracy of word sense

disambiguation phase can provide a considerable improvement on the machine translation system.

5. The neural machine translation system proposed in this research can have further improvements by the introduction of multi-sense-based word embedding instead of single-sense word embedding model. The sequence to sequence model can also be improved by using dual encoder and decoder in it, which may help in improving the learning by the network.
6. The sequence to sequence machine translation system can also be modified by the introduction of pivot language in it. This may help to improve the overall accuracy even with lesser resources.

On the whole, the future improvements on Hindi to Tamil machine translation system can be made by increasing the parallel corpus and by applying the transfer-based approach along with some other machine translation approaches. These approaches can also be extended further on the phrase level to maintain the relationship between words. The alignment phase which was used in the statistical approaches can be modified such that it can consider the global association instead of the local association of word in the sentence.

REFERENCES

[1] Papineni K., Roukos S, Ward. T, and Zhu. W, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40Th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.

[2] Goutte C., Cancedda N., Dymetman M., and Foster G., *Learning Machine Translation*. 2009.

[3] Okpor M. D., “Machine Translation Approaches: Issues and Challenges,” *Int. J. Comput. Sci. Issues*, vol. 11, no. 5/2, pp. 159–165, 2014.

[4] Choudhary N., Jha G. N., “Creating multilingual parallel corpora in Indian languages,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, pp. 527–537.

[5] Lopez A., “Statistical machine translation,” *ACM Comput. Surv.*, vol. 40, no. 3, pp. 1–49, 2008.

[6] Toral A., “Pivot-based Machine Translation between Statistical and Black Box systems,” in *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, 2012, pp. 321–328.

[7] Paul M., Yamamoto H., Sumita E., and Nakamura S., “On the Importance of Pivot Language Selection for Statistical Machine Translation,” in *NAACL ’09 Proceedings of Human Language Technologies: The 2009 Annual conference of North American Chapter of the Association for Computational Linguistics*, 2009, pp. 221–224.

[8] Saini S., “A Survey of Machine Translation Techniques and Systems for Indian Languages,” in *Proceedings of International conference on Computational Intelligence and Communication Technology (CICT)*, 2015, pp. 676–681.

[9] Dwivedi S. K., Sukhadeve P. P., “Machine Translation System in Indian Perspectives,” *J. Comput. Sci.*, vol. 6, no. 10, pp. 1082–1087, 2010.

[10] Sindhu D. V., Sagar B. M., “Study on machine translation approaches for Indian languages and their challenges,” in *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT ’16)*, 2016, pp. 262–267.

[11] Gupta D., Chatterjee N., “Identification of Divergence for English to Hindi EBMT,” in *Proceedings of the MT Summit IX*, 2003.

[12] Sinha R. M. K., Jain A., “AnglaHindi: an English to Hindi machine-aided translation system,” in *In MT Summit IX, New Orleans, Louisiana, USA*, 2003.

[13] Dhore M. L., Dixit S. K., “English to Devanagari Translation for UI Labels of Commercial

Web based Interactive Applications," *Int. J. Comput. Appl.*, vol. 35, no. 10, pp. 6–12, 2011.

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

[15] El Maazouzi Z., El Mohajir B. E., and Al Achhab M., "A technical reading in statistical and neural machines translation (SMT & NMT)," in *Proceedings of 8th International Conference on Information Technology (ICIT 2017)*, 2017, pp. 157–165.

[16] Goyal V., Lehal G. S., "Web Based Hindi to Punjabi Machine Translation System," *J. Emerg. Technol. Web Intell.*, vol. 2, pp. 148–151, 2010.

[17] Goyal V., Lehal G. S., "Hindi to Punjabi machine translation system," in *Information Systems for Indian Languages*, Springer Berlin Heidelberg, 2011, pp. 236–241.

[18] Dave S., Parikh J., and Bhattacharyya P., "Interlingua-based English – Hindi Machine Translation and Language Divergence," *Mach. Transl.*, vol. 16, no. 4, pp. 251–304, 2001.

[19] Godase A., Govilkar S., "Machine Translation Development for Indian Languages and Its Approaches," *Int. J. Nat. Lang. Comput.*, vol. 4, no. 2, pp. 55–74, 2015.

[20] Kunchukuttan A., Mishra A., Chatterjee R., Shah R., and Bhattacharyya P., "Sata-Anuvadak: Tackling Multiway Translation of Indian Languages," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

[21] Sinha R. M. K., Sivaraman K., Agrawal A., Jain R., Srivastava R., and Jain A., "ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages," in *Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century., IEEE International Conference on*, 1995, vol. 2, pp. 1609–1614.

[22] Darbari H., "Computer Assisted Translation System- An Indian Perspective Center for Development of Advanced Computing," *Mach. Transl.*, pp. 80–85, 1999.

[23] Chatterji S. and Roy D., "A Hybrid Approach for Bengali to Hindi Machine Translation T ranslation," in *Proceedings of ICON 2009: 7th International Conference on Natural Language Processing*, 2009.

[24] Ramanathan A., Bhattacharyya P., Hegde J., Shah R., and Sasikumar M., "Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation," in *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, 2008, pp. 513–520.

[25] Imam A. H., Raihan M., Arman M., Chowdhury S. H., and Mahmood K., "Impact of Corpus Size and Quality on English-Bangia Statistical Machine Translation System," in *Proceedings of 14th International Conference on Computer and Information Technology (ICCIT)*, 2011, pp. 566–571.

[26] Sánchez-Martínez F., Pérez-Ortiz J. A., and Forcada M. L., "Speeding up target-language

driven part-of-speech tagger training for machine translation," in *Advances in Artificial Intelligence, proceedings of the 5th Mexican international conference on artificial intelligence*, 2006, vol. 4293, pp. 844–854.

[27] Rabiner L. R., "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[28] Baum L. E., "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," in *Proceedings of the Third Symposium on Inequalities*, 1972, pp. 1–8.

[29] Sukhoo A., Bhattacharyya P., and Soobron M., "Translation between English and Mauritian Creole: A statistical machine translation approach," in *Proceedings of 2014 Conference and Exhibition, IST-Africa 2014*, 2014, pp. 1–10.

[30] Koehn P., Och F. J., Marcu D., "Moses: Open Source Toolkit for Statistical Machine Translation," *Proc. ACL*, no. June, pp. 177–180, 2006.

[31] Wang R., Zhao H., Lu B.-L., Utiyama M., and Sumita E., "Neural Network Based Bilingual Language Model Growing for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)*, 2014, pp. 189–195.

[32] Xiong D., Zhang M., and Wang X., "Topic-Based Coherence Modeling for Statistical Machine Translation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 3, pp. 483–493, 2015.

[33] Venkatapathy S., Bangalore S., "Three models for discriminative machine translation using Global Lexical Selection and Sentence Reconstruction," in *SSST '07 Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, 2007, pp. 96–102.

[34] Och F. J., Ney H., "A Systematic Comparison of Various Statistical Alignment Models," *J. Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, 2003.

[35] Xia Y., He. D., Qin T., Wang L., Yu N., Liu T., Ma W., "*Dual Learning for Machine Translation"*. Curran Associates, Inc., 2016.

[36] Sutskever I., Vinyals O., and Le Q. V., "Sequence to Sequence Learning with Neural Networks," in *Proceedings of Neural Information Processing Systems (NIPS 2014)*, 2014, pp. 1–9.

[37] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., and Salakhutdinov R., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[38] Shahnawaz N. A., Mishra R. B., "An English to Urdu translation model based on CBR, ANN and translation rules," *Int. J. Adv. Intell. Paradig.*, vol. 7, pp. 1–23, 2015.

[39] Bakhouche A., Yamina T., Schwab D., and Tchechmedjiev A., “Ant colony algorithm for Arabic word sense disambiguation through English lexical information,” *Int. J. Metadata, Semant. Ontol.*, vol. 10, no. 3, pp. 202–211, 2015.

[40] Lesk M., “Automatic sense disambiguation using machine readable dictionaries,” *Proc. 5th Annu. Int. Conf. Syst. Doc. - SIGDOC ’86*, pp. 24–26, 1986.

[41] Banerjee S., Pedersen T., “An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet,” in *In proceedings of third International Conference on Intelligent Text Processing and Computational Linguistics*, 2002, pp. 136–145.

[42] Soltani M., Faili H., “Target word selection in English to Persian translation using unsupervised approach,” *Int. J. Artif. Intell. Soft Comput.*, vol. 3, no. 2, pp. 125–142, 2012.

[43] Zhang B., Xiong D., Su J., and Duan H., “A Context-Aware Recurrent Encoder for Neural Machine Translation,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 12, pp. 2424–2432, 2017.

[44] Jiajun Zhang C. Z., “Deep Neural Networks in Machine Translation : An Overview,” *IEEE Intell. Syst.*, vol. 30, no. 5, pp. 16–25, 2015.

[45] Yang Z., Chen W., Wang F., and Xu B., “Multi-sense based neural machine translation,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 3491–3497.

[46] Brunning J., “Alignment Models and Algorithms for Statistical Machine Translation (PhD Thesis),” 2010.

[47] Brown P. F., Della Pietra S. A., Della Pietra V. J., and Mercer R. L., “The mathematics of statistical machine translation: Parameter estimation,” *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, 1993.

[48] Makwana M. T., Vegda D. C., “Survey:Natural Language Parsing For Indian Languages,” *Comput. Res. Repos.*, vol. abs/1501.0, 2015.

[49] Bhattacharyya P., “Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality,” *CSI J. Comput.*, vol. 1, no. 2, pp. 1–11, 2012.

[50] Gale W. A., Church K. W., “Identifying word correspondence in parallel texts,” *Proc. Fourth DARPA Work. Speech Nat. Lang. - HLT ’91*, pp. 152–157, 1991.

[51] Tiedemann J., “Word alignment - step by step,” in *Proceedings of the 12th Nordic Conference on Computational Linguistics (NODALIDA99)*, 1999, pp. 216–227.

[52] Tiedemann R., “Combining Clues for Word Alignment,” in *Proceeding of the tenth conference on European chapter of the Association for Computational Linguistics*, 2003, pp. 339–346.

[53] Srivastava J., Sanyal S., “A Hybrid Approach for Word Alignment in English-Hindi Parallel

Corpora with Scarce Resources," in *2012 International Conference on Asian Language Processing (IALP),* 2012, pp. 185–188.

[54] Joshi N., Darbari H., and Mathur I., "HMM based POS tagger for Hindi," in *Proceedings of 2nd International Conference on Artificial Intelligence and Soft Computing*, 2013, vol. 3, no. 6, pp. 341–349.

[55] Bhattacharyya P., "IndoWordNet," in *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10),* 2010, pp. 3785–3792.

[56] Pino J., Eskenazi M., "An Application of Latent Semantic Analysis to Word Sense Discrimination for Words with Related and Unrelated Meanings," in *Proceedings of the NAACL/HLT Workshop on Innovative Use of NLP for Building Educational Applications*, 2009, pp. 43–46.

[57] Katz P., Goldsmith-Pinkham P., "Word sense disambiguation using latent semantic analysis," 2006.

[58] Landauer T. K., Folt P. W., and Laham D., "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, no. 2, pp. 259–284, 1998.

[59] Kumar K. V., Yadav D., and Sharma A., "Graph Based Technique for Hindi Text Summarization," in *Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing*, 2015, vol. 339, pp. 301–310.

[60] Rosenfeld R., "Two decades of statistical language modeling: where do we go from here?," in *Proceedings of the IEEE*, 2000, vol. 88, no. 8, pp. 1270–1278.

[61] Kumar K. V., Yadav D., "Word Sense Based Hindi-Tamil Statistical Machine Translation," *Int. J. Intell. Inf. Technol.*, vol. 14, no. 1, pp. 17–27, 2018.

[62] Ahmed, Raju S. B., Chandrasekhar P. V. S., and Prasad M. K., "Application of multilayer perceptron network for tagging parts-of-speech," in *Proceedings of Language Engineering Conference, LEC 2002*, 2002, pp. 57–63.

[63] Han J., Moraga C., "The Influence of the Sigmoid Function Parameters on the Speed of Backpropagation Learning," in *Proceedings of the International Workshop on Artificial Neural Networks: From Natural to Artificial Neural Computation*, 1995, pp. 195–201.

[64] Thennarasu S., "A statistical study of tamil Morphology (Doctoral dissertation)," University of Hyderabad, 2012.

[65] Rish I., "An empirical study of the naive Bayes classifier," in *Empirical methods in artificial intelligence workshop, IJCAI*, 2001, no. JANUARY 2001, pp. 41–46.

[66] Zhang L., Marron J. S., Shen H., and Zhu Z., "Singular value decomposition and its visualization," *J. Comput. Graph. Stat.*, vol. 16, no. 4, pp. 833–854, 2007.

[67] Mantoro T., Asian J., Octavian R., and Ayu M. A., "Optimal translation of English to

Bahasa Indonesia using statistical machine translation system," in *2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, 2013, pp. 1–4.

[68] Sulaeman M. A., Purwarianti A., "Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process," in *Proceedings - 5th International Conference on Electrical Engineering and Informatics (ICEEI 2015)*, 2015, pp. 54–58.

[69] Mikolov T., Chen K., Corrado G., and Dean J., "Efficient Estimation of Word Representations in Vector Space," in *International conference on Learning Representations*, 2013, pp. 1–12.

[70] Mikolov T., Chen K., Corrado G., and Dean J., "Distributed-Representations-of-Words-and-Phrases-and-Their-Compositionality," *Comput. Res. Repos.*, vol. abs\1310.4, pp. 1–9, 2013.

[71] Medsker L. R., Jain L. C., "*Recurrent neural networks: design and applications"*. 2000.

[72] Andrychowicz M., Denil M., Gomez S., Hoffman M. W., Pfau D., Schaul T., Shillingford B., de Freitas N., "Learning to learn by gradient descent by gradient descent," in *Advances in neural information processing systems 29 (NIPS 2016)*, 2016, no. Nips, pp. 3981–89.

[73] Luong M.-T., Pham H., and Manning C. D., "Effective Approaches to Attention-based Neural Machine Translation," in *Proceedings of Empirical Methods in Natural Language Processing (EMNLP '15)*, 2015.

[74] Ambati B. R., Husain S., Jain S., Sharma D. M., and Sangal R., "Two methods to incorporate local morphosyntactic features in Hindi dependency parsing," in *Proceedings of NAACL/HLT workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2010)*, 2010, pp. 22–30.

APPENDIX A

A.1 PART-OF-SPEECH TAGSET FOR HINDI

In this research, Hindi corpus on health domain was used and it was provided by Technology Development for Indian Languages (TDIL) programme, Department of Information Technology (DIT), Ministry of Communication & IT, Government of India. The corpus provided has the part-of-speech tagging for each word in it. The standard tagset used by TDIL is as below,

Table A.1: Standard POS Tagset for Hindi language

S. NO.	CATEGORY	SUB-CATEGORY-1	SUB-CATEGORY-2	ANNOTATION
1	NOUN	COMMON NOUN		N_NN
2		PROPER NOUN		N_NNP
3		NLOC		N_NST
4	PRONOUN	PERSONAL PRONOUN		PR_PRP
5		REFLEXIVE PRONOUN		PR_PRF
6		RELATIVE PRONOUN		PR_PRL
7		RECIPROCAL PRONOUN		PR_PRC
8		WH-WORD		PR_PRQ
9		INDEFINITE PRONOUN		PR_PRI

10	DEMONSTRATIVE	DEICTIC		DM_DMD
11		RELATIVE		DM_DMR
12		WH-WORD		DM_DMQ
13		INDEFINITE		DM_DMI
14	VERB	MAIN VERB		V_VM
15		AUXILIARY		V_VAUX
16	ADJECTIVE			JJ
17	ADVERB			RB
18	POSTPOSITION			PSP
19	CONJUNCTION	COORDINATOR		CC_CCD
20		SUBORDINATOR		CC_CCS
21	PARTICLES	DEFAULT		RP_RPD
22		INTERJECTION		RP_INJ
23		INTENSIFIER		RP_INTF
24		NEGATION		RP_NEG
25	QUANTIFIERS	GENERAL		QT_QTF
26		CARDINALS		QT_QTC
27		ORDINALS		QT_QTO

28	RESIDUALS	FOREIGN WORD		RD_RDF
29		SYMBOL		RD_SYM
30		PUNCTUATION		RD_PUNC
31		UNKNOWN		RD_UNK
32		ECHOWORDS		RD_ECH

A.2 PART-OF-SPEECH TAGSET FOR TAMIL

Tamil corpus used in this research was provided by Technology Development for Indian Languages (TDIL) programme, Department of Information Technology (DIT), Ministry of Communication & IT, Government of India.Each sentence in the corpus is annotated with its part-of-speech. The standard tagset used by TDIL is as below,

Table A.2: Standard POS Tagset for Tamil language

S. NO.	CATEGORY	SUB-CATEGORY-1	SUB-CATEGORY-2	ANNOTATION
1	NOUN	COMMON NOUN		N_NN
2		PROPER NOUN		N_NNP
3		NLOC		N_NST
4	PRONOUN	PERSONAL PRONOUN		PR_PRP
5		REFLEXIVE PRONOUN		PR_PRF
6		RELATIVE PRONOUN		PR_PRL
7		RECIPROCAL PRONOUN		PR_PRC
8		WH-WORD		PR_PRQ
9	DEMONSTRATIVE	DEICTIC		DM_DMD
10		RELATIVE		DM_DMR

11		WH-WORD		DM_DMQ
12	VERB	MAIN VERB		V_VM
13			Finite	V_VM_VF
14			Non-finite	V_VM_VNF
15			Infinitive	V_VM_VINF
16			Gerund	V_VM_VNG
17		VERBAL		V_VN
18		AUXILIARY		V_VAUX
20	ADJECTIVE			JJ
21	ADVERB			RB
22	POSTPOSITION			PSP
23	CONJUNCTION	COORDINATOR		CC_CCD
24		SUBORDINATOR		CC_CCS
25			Quotative	CC_CCS_UT
26	PARTICLES	DEFAULT		RP_RPD
27		INTERJECTION		RP_INJ
28		INTENSIFIER		RP_INTF
29		NEGATION		RP_NEG

30	QUANTIFIERS	GENERAL		QT_QTF
31		CARDINALS		QT_QTC
32		ORDINALS		QT_QTO
33	RESIDUALS	FOREIGN WORD		RD_RDF
34		SYMBOL		RD_SYM
35		PUNCTUATION		RD_PUNC
36		UNKNOWN		RD_UNK
37		ECHOWORDS		RD_ECH

APPENDIX B

B.1 TRANSLITERATION USED FOR HINDI TEXT

In this thesis, Hindi letters were transliterated for better understanding using the below mentioned list,

Hindi vowels	अ	आ	इ	ई	उ	ऊ	ऋ
Roman letters	A	aa	i	ee	u	oo	R
Hindi consonants							
क	K	का (ka)	कि (ki)	की (kee)	कु (ku)	कू (koo)	कृ (kr)
ख	Kh	खा (kha)	खि (khi)	खी (khee)	खु (khu)	खू (khoo)	खृ (khr)
ग	G	गा (ga)	गि (gi)	गी (gee)	गु (gu)	गू (goo)	गृ (gr)
घ	Gh	घा (gha)	घि (ghi)	घी (ghee)	घु (ghu)	घू (ghoo)	घृ (ghr)
च	Ch	चा (cha)	चि (chi)	ची (chee)	चु (chu)	चू (cho)	चृ (chr)
छ	Chh	छा (chha)	छि (chhi)	छी (chhee)	छु (chhu)	छू (chhoo)	छृ (chhr)
ज	J	जा (ja)	जि (ji)	जी (jee)	जु (ju)	जू (joo)	जृ (jr)
झ	Jh	झा (jha)	झि (jhi)	झी (jhee)	झु (jhu)	झू (jhoo)	झृ (jhr)
ञ	N	ञा (na)	ञि (ni)	ञी (nee)	ञु (nu)	ञू (noo)	ञृ (nr)
ट	T	टा (ta)	टि (ti)	टी (tee)	टु (tu)	टू (too)	टृ (tr)
ठ	Th	ठा (tha)	ठि (thi)	ठी (thee)	ठु (thu)	ठू (thoo)	ठृ (thr)
ड	D	डा (da)	डि (di)	डी (dee)	डु (du)	डू (doo)	डृ (dr)
ढ	Dh	ढा (dha)	ढि (dhi)	ढी (dhee)	ढु (dhu)	ढू (dhoo)	ढृ (dhr)
ण	N	णा (na)	णि (ni)	णी (nee)	णु (nu)	णू (noo)	णृ (nr)
त	T	ता (ta)	ति (ti)	ती (tee)	तु (tu)	तू (too)	तृ (tr)
थ	Th	था (tha)	थि (thi)	थी (thee)	थु (thu)	थू (thoo)	थृ (thr)
द	D	दा (da)	दि (di)	दी (dee)	दु (du)	दू (doo)	दृ (dr)
ध	Dh	धा (dha)	धि (dhi)	धी (dhee)	धु (dhu)	धू (dhoo)	धृ (dhr)
न	N	ना (na)	नि (ni)	नी (nee)	नु (nu)	नू (noo)	नृ (nr)
प	P	पा (pa)	पि (pi)	पी (pee)	पु (pu)	पू (poo)	पृ (pr)
फ	Ph	फा (pha)	फि (phi)	फी (phee)	फु (phu)	फू (phoo)	फृ (phr)
ब	B	बा (ba)	बि (bi)	बी (bee)	बु (bu)	बू (boo)	बृ (br)
भ	Bh	भा (bha)	भि (bhi)	भी (bhee)	भु (bhu)	भू (bhoo)	भृ (bhr)
म	M	मा (ma)	मि (mi)	मी (mee)	मु (mu)	मू (moo)	मृ (mr)

य	Y	या (ya)	यि (yi)	यी (yee)	यु (yu)	यू (yoo)	यृ (yr)
र	R	रा (ra)	रि (ri)	री (ree)	रु (ru)	रू (roo)	रृ (rr)
ल	L	ला (la)	लि (li)	ली (lee)	लु (lu)	लू (loo)	लृ (lr)
व	V	वा (va)	वि (vi)	वी (vee)	वु (vu)	वू (voo)	वृ (vr)
श	Sh	शा (sha)	शि (shi)	शी (shee)	शु (shu)	शू (shoo)	शृ (shr)
ष	Sh	षा (sha)	षि (shi)	षी (shee)	षु (shu)	षू (shoo)	षृ (shr)
स	S	सा (sa)	सि (si)	सी (see)	सु (su)	सू (suoo)	सृ (sr)
ह	H	हा (ha)	हि (hi)	ही (hee)	हु (hu)	हू (hoo)	हृ (hr)

Hindi vowels	ए	ऐ	ओ	औ	अं	अः
Roman letters	E	ai	o	au	an	ah
Hindi consonants						
क	के (ke)	कै (kai)	को (ko)	कौ (kau)	कं (kan)	कः (kah)
ख	खे (khe)	खै (khai)	खो (kho)	खौ (khau)	खं (khan)	खः (khah)
ग	गे (ge)	गै (gai)	गो (go)	गौ (gau)	गं (gan)	गः (gah)
घ	घे (ghe)	घै (ghai)	घो (gho)	घौ (ghau)	घं (ghan)	घः (ghah)
च	चे (che)	चै (chai)	चो (cho)	चौ (chau)	चं (chan)	चः (chah)
छ	छे (chhe)	छै (chhai)	छो (chho)	छौ (chhau)	छं (chhan)	छः (chhah)
ज	जे (je)	जै (jai)	जो (jo)	जौ (jau)	जं (jan)	जः (jah)
झ	झे (jhe)	झै (jhai)	झो (jho)	झौ (jhau)	झं (jhan)	झः (jhah)
ञ	ञे (ne)	ञै (nai)	ञो (no)	ञौ (nau)	ञं (nan)	ञः (nah)
ट	टे (te)	टै (tai)	टो (to)	टौ (tau)	टं (tan)	टः (tah)
ठ	ठे (the)	ठै (thai)	ठो (tho)	ठौ (thau)	ठं (than)	ठः (thah)
ड	डे (de)	डै (dai)	डो (do)	डौ (dau)	डं (dan)	डः (dah)
ढ	ढे (dhe)	ढै (dhai)	ढो (dho)	ढौ (dhau)	ढं (dhan)	ढः (dhah)
ण	णे (ne)	णै (nai)	णो (no)	णौ (nau)	णं (nan)	णः (nah)
त	ते (te)	तै (tai)	तो (to)	तौ (tau)	तं (tan)	तः (tah)
थ	थे (the)	थै (thai)	थो (tho)	थौ (thau)	थं (than)	थः (thah)
द	दे (de)	दै (dai)	दो (do)	दौ (dau)	दं (dan)	दः (dah)
ध	धे (dhe)	धै (dhai)	धो (dho)	धौ (dhau)	धं (dhan)	धः (dhah)
न	ने (ne)	नै (nai)	नो (no)	नौ (nau)	नं (nan)	नः (nah)
प	पे (pe)	पै (pai)	पो (po)	पौ (pau)	पं (pan)	पः (pah)
फ	फे (phe)	फै (phai)	फो (pho)	फौ (phau)	फं (phan)	फः (phah)
ब	बे (be)	बै (bai)	बो (bo)	बौ (bau)	बं (ban)	बः (bah)
भ	भे (bhe)	भै (bhai)	भो (bho)	भौ (bhau)	भं (bhan)	भः (bhah)

म	मे (me)	मै (mai)	मो (mo)	मौ (mau)	मं (man)	मः (mah)
य	ये (ye)	यै (yai)	यो (yo)	यौ (yau)	यं (yan)	यः (yah)
र	रे (re)	रै (rai)	रो (ro)	रौ (rau)	रं (ran)	रः (rah)
ल	ले (le)	लै (lai)	लो (lo)	लौ (lau)	लं (lan)	लः (lah)
व	वे (ve)	वै (vai)	वो (vo)	वौ (vau)	वं (van)	वः (vah)
श	शे (she)	शै (shai)	शो (sho)	शौ (shau)	शं (shan)	शः (shah)
ष	षे (she)	षै (shai)	षो (sho)	षौ (shau)	षं (shan)	षः (shah)
स	से (se)	सै (sai)	सो (so)	सौ (sau)	सं (san)	सः (sah)
ह	हे (he)	है (hai)	हो (ho)	हौ (hau)	हं (han)	हः (hah)

B.2 TRANSLITERATION USED FOR TAMIL TEXT

Tamil letters were also transliterated for better understanding using the below mentioned list,

Tamil Vowels	அ	ஆ	இ	ஈ	உ	ஊ
Roman Letters	A	Ā	I	ī	U	ū
Tamil Consonants						
க	Ka	கா (kā)	கி (ki)	கீ (kī)	கு (ku)	கூ (kū)
ங	ṅa	ஙா (ṅā)	ஙி (ṅi)	ஙீ (ṅī)	ஙு (ṅu)	ஙூ (ṅū)
ச	Ca	சா (cā)	சி (ca)	சீ (cī)	சு (cu)	சூ(cū)
ஞ	Ña	ஞா (ñā)	ஞி (ñi)	ஞீ (ñī)	ஞு (ñu)	ஞூ (ñū)
ட	ṭa	டா (ṭā)	டி(ṭi)	டீ (ṭī)	டு (ṭu)	டூ (ṭū)
ண	ṇa	ணா (ṇā)	ணி (ṇi)	ணீ (ṇī)	ணு (ṇu)	ணூ (ṇū)
த	Ta	தா (tā)	தி (ti)	தீ (tī)	து (tu)	தூ (tū)
ந	Na	நா (nā)	நி (na)	நீ (nī)	நு (nu)	நூ (nū)
ப	Pa	பா (pā)	பி (pi)	பீ (pī)	பு (pu)	பூ (pū)
ம	ma	மா (mā)	மி (mi)	மீ (mī)	மு (mu)	மூ (mū)
ய	Ya	யா (yā)	யி (yi)	யீ (yī)	யு (yu)	யூ (yū)
ர	Ra	ரா (rā)	ரி (ri)	ரீ (rī)	ரு (ru)	ரூ (rū)
ல	La	லா (lā)	லி (li)	லீ (lī)	லு (lu)	லூ (lū)
வ	Va	வா (vā)	வி (vi)	வீ (vī)	வு (vu)	வூ (vū)
ழ	ḻa	ழா (ḻā)	ழி (ḻi)	ழீ (ḻī)	ழு (ḻu)	ழூ (ḻū)
ள	ḷa	ளா (ḷā)	ளி (ḷi)	ளீ (ḷī)	ளு (ḷu)	ளூ (ḷū)
ற	ṟa	றா (ṟā)	றி (ṟi)	றீ (ṟī)	று (ṟu)	றூ (ṟū)
ன	ṉa	னா (ṉā)	னி (ṉi)	னீ (ṉī)	னு (ṉu)	னூ (ṉū)

Tamil Vowels	எ	ஏ	ஐ	ஒ	ஓ	ஔ	ஃ
Roman Letters	e	Ē	Ai	o	Ō	oḷa	ḥ
Tamil Consonants							
க	கெ (ke)	கே (kē)	கை (kai)	கொ (ko)	கோ (kō)	கௌ (keḷa)	க் (kḥ)
ங	ஙெ (ṅe)	ஙே (ṅē)	ஙை (ṅai)	ஙொ (ṅo)	ஙோ (ṅō)	ஙௌ (ṅeḷa)	ங் (ṅḥ)
ச	செ (ce)	சே (cē)	சை (cai)	சொ (co)	சோ (cō)	சௌ (ceḷa)	ச் (cḥ)
ஞ	ஞெ (ñe)	ஞே (ñē)	ஞை (ñai)	ஞொ (ño)	ஞோ (ñō)	ஞௌ (ñeḷa)	ஞ் (ñḥ)
ட	டெ (ṭe)	டே (ṭē)	டை (ṭai)	டொ (ṭo)	டோ (ṭō)	டௌ (ṭeḷa)	ட் (ṭḥ)
ண	ணெ (ṇe)	ணே (ṇē)	ணை (ṇai)	ணொ (ṇo)	ணோ (ṇō)	ணௌ (ṇeḷa)	ண் (ṇḥ)
த	தெ (te)	தே (tē)	தை (tai)	தொ (to)	தோ (tō)	தௌ (teḷa)	த் (tḥ)
ந	நெ (ne)	நே (nē)	நை (nai)	நொ (no)	நோ (nō)	நௌ (neḷa)	ந் (nḥ)
ப	பெ (pe)	பே (pē)	பை (pai)	பொ (po)	போ (pō)	பௌ (peḷa)	ப் (pḥ)
ம	மெ (me)	மே (mē)	மை (mai)	மொ (mo)	மோ (mō)	மௌ (meḷa)	ம் (mḥ)
ய	யெ (ye)	யே (yē)	யை (yai)	யொ (yo)	யோ (yō)	யௌ (yeḷa)	ய் (yḥ)
ர	ரெ (re)	ரே (rē)	ரை (rai)	ரொ (ro)	ரோ (rō)	ரௌ (reḷa)	ர் (rḥ)
ல	லெ (le)	லே (lē)	லை (lai)	லொ (lo)	லோ (lō)	லௌ (leḷa)	ல் (lḥ)
வ	வெ (ve)	வே (vē)	வை (vai)	வொ (vo)	வோ (vō)	வௌ (veḷa)	வ் (vḥ)
ழ	ழெ (ḻe)	ழே (ḻē)	ழை (ḻai)	ழொ (ḻo)	ழோ (ḻō)	ழௌ (ḻeḷa)	ழ் (ḻḥ)
ள	ளெ (ḷe)	ளே (ḷē)	ளை (ḷai)	ளொ (ḷo)	ளோ (ḷō)	ளௌ (ḷeḷa)	ள் (ḷḥ)
ற	றெ (ṟe)	றே (ṟē)	றை (ṟai)	றொ (ṟo)	றோ (ṟō)	றௌ (ṟeḷa)	ற் (ṟḥ)
ன	னெ (ṉe)	னே (ṉē)	னை (ṉai)	னொ (ṉo)	னோ (ṉō)	னௌ (ṉeḷa)	ன் (ṉḥ)

LIST OF AUTHOR'S PUBLICATIONS

International Journals

1. Vimal Kumar. K, Divakar Yadav, 2018: Word Sense Based Approach for Hindi to Tamil Machine Translation Using English as Pivot Language. *Journal on Advanced Intelligence Paradigms*, doi: 10.1504/IJAIP.2018.10008778. (h-index [**6**], h5-index **[7]**, h5-median [**8**]), SJR=0.199, ***(serial no. 2498 in UGC list of Journals).***

   Indexed in: *Academic OneFile, ACM Digital Library, cnpLINKer, DBLP Computer Science Bibliography, Education Research Abstracts, Expanded Academic ASAP, Google Scholar, Info Trac, INspec (Institution of Engineering & Technology) SCImago,* ***SCOPUS****.*

2. Vimal Kumar. K, Divakar Yadav, 2018: Word Sense Based Hindi-Tamil Statistical Machine Translation. *Journal on Intelligent Information Technologies*, doi: 10.4018/IJIIT.2018010102. (h-index: [10]), SJR=0.239, ***(serial no. 22986 in UGC list of Journals).***

   Indexed in: *ACM Digital Library, Australian Business Deans Council (ABDC), Burrelle's Media Directory, Cabell's Directories, CSA Illumina, DBLP, DEST Register of Refereed Journals, Directory of publications & Broadcast media, GetCited, Google Scholar, Journal TOCs, Library & Information Science Abstracts, MediaFinder,Norwegian Social Data Services (NSD), The index of Information systems Journals, The standard Periodical directory, Ulrich's Periodicals Directory, Web of Science, Web of Science Emerging Sources Citation Index (ESCI),* ***SCOPUS****, Compendex (Elsevier Engineering Index), INSPEC, SCIMago*

International Conferences

1. Vimal Kumar. K, Divakar Yadav, 2015: An improvised extractive approach to Hindi text summarization, *International conference on Information System Design & Intelligent Applications*, Advances in Intelligent and soft computing (AISC) – Springer, Vol. 339,

ISBN: 978-81-322-2249-1, ISSN: 2194-5357, pp. 291-300, 8-9 Jan 2015. doi:10.1007/978-81-322-2250-7_28. **(Cited by 6), Indexed in ISI Proceedings, EL-Compendex, DBLP, SCOPUS, Google Scholar and SpringerLink.**

2. Vimal Kumar. K, Divakar Yadav, Arun Sharma, 2015: A Graph based technique to Hindi text summarization, *International conference on Information System Design & Intelligent Applications*, Advances in Intelligent and soft computing (AISC) – Springer, Vol. 339, ISBN: 978-81-322-2249-1, ISSN: 2194-5357, pp. 301-310, 8-9 Jan 2015. doi:10.1007/978-81-322-2250-7_29. **(Cited by 4), Indexed in ISI Proceedings, EL-Compendex, DBLP, SCOPUS, Google Scholar and SpringerLink.**
3. Vimal Kumar. K, Yamuna Prasad, 2019: A sequential approach to handle machine translation of low resource languages, *In the proceedings of 20th International conference on Computational linguistics & Intelligent text processing (CICLING), pp XX-XX, 2019,* **Indexed in SCOPUS, DBLP, Web of Science.** (h5-index [**21**], h5-median [**33**]).

A Novel Approach for Sense based Hindi to Tamil Machine Translation System

1. Introduction

Machine translation is the process of translating source text to the destination text using the features of both the languages. Mapping the features of two different languages is not an easy task which generally needs clear understanding of both languages. Thus, machine translation requires the knowledge of both the languages. There are various existing approaches for machine translation – rule based machine translation, statistical machine translation, example based machine translation and hybrid machine translation. Based on the literature survey, it is found that translation system that exists for the Hindi to Tamil has very less accuracy. The Hindi language is a partially free word ordered whereas Tamil is fully free word ordered language. The accuracy of these translation systems can further be improved by various methods and thus the research is narrowed to the Hybrid Machine Translation approach. The main objective of this research is to develop a Hindi-Tamil machine translation system by considering the syntactic and semantic features of both the languages. For this purpose a hybrid approach is used, which makes use of transfer based machine translation which is coupled with statistical machine translation for post processing. These natural languages have specific unique features which distinguishes it from each other. In this transfer based machine translation system, there is a need for an intermediate language/representation, which can be natural language by itself, called as pivot language. This translation mechanism requires the understanding of the source language which will be converted to an intermediate language. Based on the target language information, the intermediate language is converted to the target language. Before the translation process, there is need for source language analysis too, which is made using morphological analysis and lexical categorization. Morphological analysis will analyze the surface form of the source language and output the part-of-speech along with the sub-category information. Lexical categorization is used to reduce the ambiguous meaning of every word in the sentence by considering the contextual meaning of the word. Based on the analysis, the statistical approach is applied to perform the translation to the intermediate pivot language and this method is known as Lexical Transfer. This lexical transfer translates the languages word-by-word which is subjected to structural transfer in order to make it grammatically correct. During the structural transfer, the intermediate language is grammatically aligned to the pivot language using the statistical information. The translated pivot language is further converted to the target language using statistical approach.

There are words in source language that have different translations in target language and this can be identified using word sense disambiguation over the sentence. This word sense disambiguation will provide the contextual information of the word under consideration. The contextual information can provide an accurate target word. Thus, the objective of this system is to improve the efficiency of the statistical machine translation by providing additional information such as the part-of-speech of the word to be translated as well as the information about the preceding words and to make use of the word's sense in the translation process. Since this translation process takes in to account the word's part-of-speech (syntactic information) as well as the word's sense (semantic information) during the machine translation, it improves the efficiency and accuracy of the translation.

Based on the research by Paul, M. et. al. [1], the English language is identified as pivot language due to its vast resource availability. The mapping between Indian languages and the pivot language (English) is also a bottleneck in this research as the grammar for these languages are totally different. Moreover, English is a fixed word order language whereas Indian languages (Hindi and Tamil) are free word order languages. To overcome this issue, the words in both languages has to be mapped and aligned to each other. There is an existing word alignment algorithm known as IBM model on word alignment. These IBM models never include the part-of-speech of the word to find the alignment parameters. Based on the analysis, it is found that the words alignment parameters changes with respect to its part-of-speech. Thus in this proposed system, we introduce a modified word alignment algorithm which considers the part-of-speech of the word too. Using the aligned words, the proposed system performs a transfer from Hindi to English language using Bayesian method. But the performance of Bayesian method degrades the systems accuracy as there is more semantic distortion because of the pivot language, which can be overcome by the introduction of semantic analysis to interpret the words sense in the input language. The same Bayesian method is used for the transfer from English to Tamil as well. The overall performance of translation system that includes word sense disambiguation and pivot language is more efficient than the one without word sense disambiguation.

To improve the accuracy further, a sequence to sequence model has been developed and this model makes use of Long-Short term memory (LSTM) neural network to capture the mapping between two languages under consideration. The semantics in each of the language is being captured using word2vector model. The word2vector can be generated using continuous bag-of-words model. These generated vectors are further used for training the sequence to sequence model, so that, it can map the relations between the sentences.

2. Motivation

Based on the literature survey, the following issues were identified which motivated to propose the solution for this research,

i. Machine translation system for Indian language pairs has very less accuracy
ii. The existing systems do not consider the contextual information of the words during translation
iii. In general, the resources for Indian languages are very poor
iv. As compared with English, Hindi and Tamil are free word order languages.
v. Semantics are lost during translation from Hindi to Tamil using English as a pivot.

Issue 1: Machine translation system for Indian language pairs has very less accuracy

Indian languages are free word order languages and morphological structure of words in these languages are different as compared to English. The existing machine translation approach for English language won't be suffice for the Indian languages. Morphological structure of words in Hindi and Tamil has information such as TAM (tense, aspect and modality). This TAM information plays a vital role during translation process. This TAM information can be detected using morphological analysis and tagger methods. But the accuracy of these methods is still poor compared to other languages. Thus, the accuracy of translation system also degrades with respect to decrease in accuracy of morphological analysis and tagger in Indian languages.

Issue 2: The existing systems do not consider the contextual information of the words during translation

In Hindi and Tamil languages, sense of the word changes according to the context where it is being used. For example, the word 'Aam' has different sense in both these phrases 'Aam aadhmi' and 'Aam ka ped'. In case the contextual information is not used during translation, the meaning of the translated sentence may get distorted and the actual information which was conveyed in the source text won't be available in the target text. Thus, the contextual information plays a vital role too.

Issue 3: In general, the resources for Indian languages are very poor

The resource availability for Indian languages are still poor compared to English and other European languages. Technology development for Indian languages (TDIL) has taken initiative to develop the various resources for Indian languages [18]. But still the corpus availability is poor which has vital impact on the accuracy of the system.

Issue 4: As compared with English, Hindi and Tamil are free word order languages

English has a fixed grammatical sequence as compared with Indian languages. Since Hindi and Tamil are free word order languages, the sentences in these languages are grammatically valid irrespective of the order of words. The existing state-of-art methods accuracy is very low due to free word nature of Indian languages. When the size of corpus is increased, the system generates more number of possible translation for the input sentence. Thus, it degrades the accuracy of the system.

Issue 5: Semantics are lost during translation from Hindi to Tamil using English as a pivot

The semantic information that is being conveyed in the source text (Hindi) will be lost in case if English is used as an intermediate language for the translation to Tamil. Certain words in Hindi has a one-to-one mapping with the respective Tamil words whereas the mapping with English has a many to one or one to many relationships. Thus, semantics are lost during the translation to the pivot language.

3. Objective

To develop a word sense Hindi to Tamil machine translation system to fulfill the research gap that has been identified. Thus, the proposed objective of this research is defined as follows,

i. The accuracy of the machine translation can be improved by using both the syntactic and semantic information during the translation process.

ii. Since there is resource scarcity, use of a pivot language will be more helpful in improving the translation process. The identified pivot language is English, which has abundant resources.

iii. Use of word sense disambiguation can provide the information required for context based translation

iv. To improve the accuracy for translation further, a neural machine translation has been developed which is a sequence to sequence model for Hindi to Tamil machine translation.

v. The semantic information is retained using a neural machine translation such that the information conveyed in source text is not lost during the translation.

4. Related Work

In Sutskever et. al [21], sequence to sequence learning with neural networks was developed, which consists of an encoder and decoder. Encoder converts the input sentence into a fixed-length vector. These vectors are further converted into target sentences by the decoder module. Both these encoder and decoder are jointly trained such that the probability of correct translation is maximized.

A simple way to prevent overfitting in neural network was introduced by Srivastava et. al [22]. In this paper, the author introduced dropout regularization. During the training of neural network, the weights of neuron are randomly made zero based on the dropout percentage which prevents the neurons from adapting too much according to the training data. Author has also found that it improves the problem of overfitting in significant manner as compared with the other regularization methods.

Ali Hasan Imam et al has discussed about the impact of corpus size in English-Bangla statistical machine translation [13]. The author has identified that increase in corpus size to improve translation quality will saturate the quality of translation at particular instant and then, there comes the need for improving quality of the corpus. Author has developed their own corpus from various sources and has evaluated the machine translation system based on BLEU score.

Vishal Goyal and Dr. G. S. Josan proposed a machine translation system mainly from Hindi-Punjabi in 2009 at Punjabi University Patiala. It is based on direct word-to-word translation and reported 95% accuracy [7]. In 2010, Vishal Goyal and Prof. G.S. Lehal proposed a machine translator from Hindi to Punjabi using direct translation and later on improving the language learning modules for the enhancement of the quality of the system. The accuracy of the translation is approximately 95%. [8].

In 2014, the research on statistical machine translation between English and Mauritian Creole language pairs has been developed specifically for tourism and business purpose [14]. The author has used MOSES tool to develop the system and found that system performance was not up to the mark as the parallel corpus was too small. Author has also used the bilingual dictionary to improve its performance and based on the system's evaluation they has found that BLEU score to be approximately 6.0. The BLEU score seems to increase with respect to the increase in corpus size.

Rui Wang et al have proposed a novel neural network bilingual language model for statistical machine translation system [15]. The continuous space language model makes use of monolingual corpus and the author has proposed a method to modify it to bilingual continuous space language model. They have used the different language models over Chinese-English machine translation system and have used 1 million parallel training data. Author has also suggested the way to reduce the computational and space complexity in processing the identified phrases by considering only top phrases which are identified using ranking method. This proposed system is found to outperform the existing language models converting/growing methods in the statistical machine translation systems.

In 2015, DeyiXiong et al has proposed a document level topic-based coherence model for statistical machine translation [16]. This system extracts the coherence chain from the source text and this extracted chain is further mapped on to the target coherence chain using maximum entropy classifier. Author has developed two topic-based coherence models over these generated target coherence chain – word level coherence model and the phrase level coherence model. The author has found that the developed system outperforms compared to the baseline system. It is also identified that the phrase level coherence model is comparatively better over word level coherence model.

Annotation in Hindi and English varies highly as both the languages differ in terms of grammar. The two techniques for tagging can be: linking all possible words in both source and target language, or, linking least possible words in both source and the target language. Since exact correspondence is hard to find and word to word translation is possible for very few words, thus remaining words are translated based on the combination of words. Thus annotation can follow certain guidelines such as following fuzzy, regular and null links in punctuation which can lead to desired results [9].

For English to Hindi statistical machine translation, the key challenge is that the Hindi language is richer in morphology than the English language. Reordering of English source sentences in accordance with Hindi language and/ or making use of suffixes of Hindi words are the two strategies that can facilitate reasonable performance. Since Indian languages and English differ in word order, morphologically rich Indian languages and unavailability of huge parallel corpora for two languages make the above two strategies more challenging to derive the desired results with reasonable performance[11].

5. Work done

The machine translation system basically decodes the target text from the source text. To translate the source to target language, there is need for three major phases in it – Source language analysis (Morphological analysis and lexical categorization), Source-Target transfer (Lexical transfer) and Target language rearrangements (structural transfer). In source language analysis phase, the system analyzes the input text to capture the syntactic and semantics features of the input text. These features are used during the lexical transfer phase to convert the source text to target language text. The transfer phase can be developed using any approach such as statistical, rule based, example based. The output of this transfer phase won't be grammatically correct as per the target language features. The last phase structural transfer is used to make the target text grammatically correct.

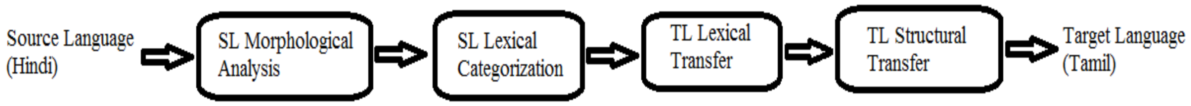

Figure 1: Overall System Architecture

The system's accuracy can be improved by using deep learning in place of statistical machine translation. The proposed system is developed using sequence to sequence neural machine translation. Since, the sequence to sequence model makes use of a special type of recurrent neural network, long short time memory network (LSTM), it captures the semantics based on the context in which it is being used.

5.1. Hindi to Tamil Statistical Machine translation using Word Sense Disambiguation

In a Hindi to Tamil statistical machine translation system, analysis phase is required for the source language (Hindi), which gives information about morphology and part-of-speech of the words. In this research source language is Hindi & target language is Tamil. The Hindi morphological analyzer separates the suffix from a word using longest suffix matching method. Suffixes provide the information about the case, the number, and the gender which are used in inflected words. To identify the word's sense in a sentence there is need for syntactic information and semantic information of the word being used. The syntactic information is identified using a multi-layered perceptron based POS tagger and this process is called as Hindi lexical categorization. In the features multilayer perceptron based tagger, the identified features for the tagging purpose are - the probability of the word given the word's tag $P(word_i/tag_i)$ and the probability of the previous word's tag given the word's tag $P(tag_{i-1}/tag_i)$. Based on the tagged corpus, these probability combinations are found and are used for the training of the multilayer preceptron. The multilayer perceptron is as shown in figure-2 below,

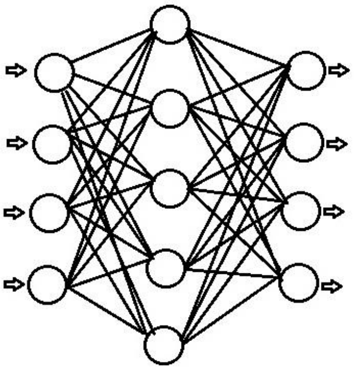

Figure 2: Multilayer perceptron based part-of-speech tagger

The hidden layer will be activate using the sigmoid function which is represented as

$$y(f(x)) = \frac{1}{1+e^{-f(x)}}$$

$$where,$$

$$f(x) = \sum_{j=1}^{N} w_j * P(word/tag_i) * P(prev\_tag_j/tag_i)$$

Thus using this multilayered perceptron, there will be an output matrix $[y_1 \quad y_2 \quad y_3 \quad \dots \quad \dots \quad \dots \quad y_N]$ where the dimension 'N' is equivalent to the number of distinct tags. The maximum value from this output matrix will used to locate the probable tag of the particular word.

To capture the semantic that is required for contextual information, word's sense disambiguation is required which is identified using Latent Semantic Analysis (LSA) and Hindi Wordnet. The semantic information of the word is identified using the Hindi Wordnet dictionary to identify the various senses for the word with its identified part-of-speech. Semantic analysis between the words under consideration over the various senses of the word provides a score of its relevance. The high scored senses are used further in the translation. LSA technique is a singular value decomposition method which decomposes the term-frequency matrix in to three different matrices namely left singular matrix (L), right singular matrix (R) and singular diagonal matrix (D) as represented in the equation below,

$$\begin{bmatrix} f_{t1}^{1} & f_{t2}^{1} & \ldots\ldots & f_{tn}^{1} \\ f_{t1}^{2} & f_{t2}^{2} & \ldots\ldots. & f_{tn}^{2} \\ \ldots & \ldots & \ldots & \ldots \\ f_{t1}^{m} & f_{t2}^{m} & \ldots\ldots & f_{tn}^{m} \end{bmatrix} = L * R * S$$

where, f_{tn}^{m} - indicates the frequency of nth-term in mth-sentence

L - Left singular matrix (called as Term matrix)

R - Right singular matrix (called as concept matrix)

S - Singular diagonal matrix (called as concept by document matrix)

After the decomposition of this matrix, the product of the right singular matrix and singular diagonal matrix gives the frequency of semantically similar words in different sentences. The similarity between two sentences is found using cosine similarity over the product of right singular vector and singular diagonal vector of the sentence considered.

At this phase, the system has identified the relevant senses for the word, now the system will be subjected to lexical transfer using Bayesian approach, which is mathematically expressed as,

$$P\left(W_t/(W_s, Tag_s)\right) = P\left((W_s, Tag_s)/W_t\right) * P\left(W_t\right)$$

Since the Bayesian method requires the proper sequence in the target language, the structural rearrangement of the sentence is required before lexical transfer. Based on the analysis, it is found that Hindi language follows Subject Object Verb (SOV) sequence that is same as in Tamil. The transfer phase makes use of the target language model and considers the preceding word information in the target language level. Accordingly, the grammatical restructuring of the words in source language happens such that the source language sentence is aligned with the grammar as in the target language. This is achieved by applying the statistical based alignment algorithm. The Hindi to Tamil machine translation system is found to perform well but, the accuracy of the system is still lagging due to the lesser resources that exist for Indian languages. Thus, there is need for a system which can overcome low-resource availability issue.

5.2. Hindi to Tamil Machine translation using pivot language

To overcome, the low-resource availability, a pivot language has been introduced in between the source and target languages which can help to improve the accuracy of the system. Pivot language based machine translation has three major modules – source language analysis, lexical transfer (source to pivot followed by pivot to target) and structural transfer. Since there is scarcity of resources there is need for pivot language which has rich resource. Thus, English is chosen as a pivot language for this pivot based approach.

The source language analysis is as same as in the previous approach. The lexical information identified in the analysis phase are used in lexical transfer phase to convert Hindi text to its respective English text using statistical approach. During the research it is found that the grammar used in the Hindi and Tamil matches in most of the cases. There won’t be any need to rearrange the sentences based on the target language grammar. Thus the lexical transfer phase translates the English language to Tamil language using the statistical information skipping pivot language structural transfer.

The proposed lexical transfer phase of Hindi-English makes use of word by word statistics based transfer. The method makes use of bilingual corpus so that the system requires less human effort to perform the lexical transfer. The corpora used in this system are on health domain. Statistical transfer takes the most probable translation in the pivot language for that input language. The statistical model of language translation finds every individual word of the source language sentence in a list and substitutes this word with the equivalent pivot language word. But this method does not ensure accurate results as the relationship followed between words in source language and in pivot language are often one-to-many, instead of one-to-one. This is due to a word in source language can give different equivalent translations in the pivot language. One of the solutions for increasing the accuracy of the results can be words being distinguished based on its part-of-speech. Thus, this statistical system for translation is developed considering the probability of target word given the source word and its corresponding part-of-speech. The probability can be represented as mentioned below,

$$P\left(\frac{w_t}{w_s, pos}\right) = P\left(\frac{w_p}{w_s, pos}\right) * P\left(\frac{w_t}{w_p, pos}\right)$$

The translation model is developed to provide the translation from one language to another, based on the analysis of both the languages. Since the translation process depends on the grammatical sequence of words in the source language, an analysis of grammatical structure of the source language is performed and a language model is developed for the same. Basically, this system makes use of MGIZA++ (a tool for word alignment). This tool generates the alignment table which is required for the statistical method. The sample alignment table is as shown in table-1 which has the position in target language, source word position, length of Hindi sentence and length of its corresponding Tamil word.

Table 1: Sample Alignment Table

S.No	**J (Word's Position in target language)**	**I (Word's Position in Source language)**	**L (Source Sentence Length)**	**M (Target Sentence Length)**	**Frequency**
1	4	3	8	10	252
2	7	7	10	10	231

But by performing the Hindi-English-Tamil (without WSD) statistical machine translation, it is found that certain information is getting distorted and changes the inner meaning of the sentence. Thus, there is a need for word's sense disambiguation in the complete Hindi to Tamil translation process.

5.3. Hindi to Tamil Machine translation using Hybrid approach

In this approach the system makes use of pivot language as well as a word sense disambiguation so that the meaning which is lost during translation via a pivot language can be preserved. The preprocessing phase is basically a source language analysis phase which is used to capture the word level information such as TAM (tense, aspect and modality). After the preprocessing, the syntactic information in the input text are captured, but, the semantics also contribute for the accuracy of the machine translation. The semantic information in the input text are identified using the word sense disambiguation phase, which makes use of latent semantic analysis (LSA). Once the word's sense is identified, the statistical method is used to transfer the Hindi text to the pivot language, English. For the statistical transfer, the system needs to be grammatically aligned with the English language, which is performed in the alignment phase. The statistical method is applied once again on transfer phase from English to Tamil.

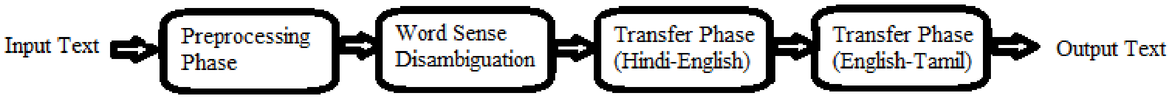
Input Text
Preprocessing Phase
Word Sense Disambiguation
Transfer Phase (Hindi-English)
Transfer Phase (English-Tamil)
Output Text

Figure 3: Word sense based Hindi-Tamil Machine translation using pivot language

Using this hybrid machine translation system, the translation won’t be capturing the TAM (tense, aspect, and modality) information stored in the input text. This information can improve the accuracy of translation further which can be captured using sequence to sequence learning technique.

5.4. A deep learning approach for Hindi to Tamil Machine translation

The sequence to sequence machine translation model is basically a combination of an encoder which encodes the input sentence into a fixed length vector and a decoder which decodes the target sentence from the fixed length vector generated in the encoding process. The encoder and decoder are trained in parallel using the vectors of languages under consideration. To train the sequence to sequence model, the vectors are generated using word2vec – continuous bag-of-words model. Continuous bag-of-words model is a deep learning model which predicts the word at a context by considering the neighboring words in the training corpus. The neighboring words before that word and after the word are considered during training of the network. These vectors which are generated for the languages under consideration, i.e., Hindi and Tamil, are further used for training the sequence to sequence model. The vectors basically represent the words in a continuous vector space and the words which are semantically similar are placed at the neighboring points in the vector space. Since the proposed system works based on syntactic and semantic features of the languages, there is a need for a method to learn this word embedding from the textual data. The vectors that are generated from word embedding are found to contain the semantic relations between the words. One such approach is the continuous bag-of-words (CBOW) model, which predicts the word at n-position if its preceding words are known to the model. In the CBOW model, a neural network is being used to learn based on the training data provided to it. The semantic information is represented in the form of vectors. During the training and extraction phase, the CBOW model considers a trigram model as it is found to be more accurate as compared with any n-gram. Thus, the CBOW model considers the two words and predicts the third one during training. Based on the training, it generates the vectors, which are further fed to the sequence to sequence based machine translation.

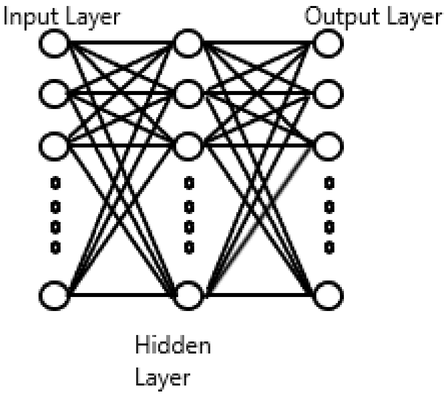

Figure 4: Continuous Bag-of-Words model for word embedding

The sequence to sequence model has a Long-Short term memory (LSTM) network at the back end. LSTM network is special variant of recurrent neural network which is capable of learning long term dependencies. This network keeps a memory of information over a long period of time so that the dependencies between them are captured properly. In machine translation, the meaning of a sentence depends on its preceding sentences in the text corpus. Thus, a LSTM is being used in this sequence to sequence model.

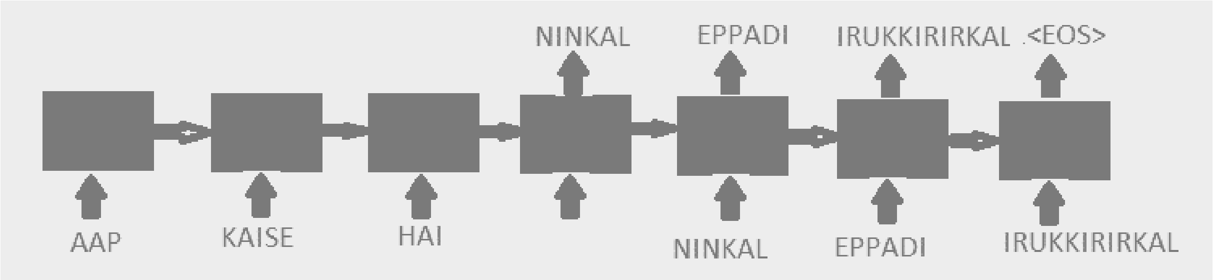

Figure 5: Architecture of Sequence to Sequence based Hindi to Tamil Machine Translation

The encoder encodes the input vector to an intermediate vector using a long short term memory (LSTM) network. Since the LSTM network needs to keep a memory of the vectors, there is LSTMCell being used in it to store these values. To protect and control these values there are four different gates being used in it – input gate, out gate, cell gate and forget gate. Based on these gate values, the LSTM network discards a value during training or keeps it during training. The input gate and cell gate are the one which decides what value is to be stored in the LSTMCell. Forget gate is used to predict whether the value must be retained for further use or not. The out gate is the used to decide upon the value that needs to be updated on the LSTMCell. Using these gate values, the LSTM network computes the output of the hidden state h_t at time *t* based on the expression,

$$h_t = o_t \tanh c_t$$

where, o_t- out gate and c_t- value at LSTMCell

As per the expression above, the output of hidden state depends on the current value in the LSTMCell and the out gate. The current value of the LSTMCell c_t is calculated using the previous cell value and the gate value based on the equation [20],

$$c_t = f_t * c_{t-1} + i_t * g_t$$

Where, f_t – forget gate and i_t- input

During the calculation of the current value of LSTMCell, the values of forget gate, input, cell and previous cell value are multiplied element-by-element. The out gate value is calculated based on the weights assigned between input and output along with the weights that are assigned between hidden and output layers. Apart from these two parameters, the out gate also uses the value of hidden state in the previous layer x_t and hidden layer value at time t-1, h_{t-1}. Sigmoid function is being used to keep the values in the range of 0 to 1. It is mathematically expressed as,

$$o_t = sigmoid(w_i^o * x_t + b_i^o + w_h^o * h_{t-1} + b_h^o)$$

Where,

w_i^o- Weight vector between input layer and out gate

w_h^o- Weight vector between hidden layer and out gate

b_h^o- Bias between hidden layer and out gate

b_i^o- Bias between input layer and out gate

The values of LSTMCell and the out gate value are the primary requirement of the LSTM network to predict the vector at the hidden state h_t. But these two values, LSTMCell value and out gate value, are dependent on other parameters such as, input gate, forget gate and cell gate, which are further calculated using the below mentioned equations,

$$g_t = \tanh(w_i^g * x_t + b_i^g + w_h^g * h_{t-1} + b_h^g)$$

$$f_t = sigmoid\left(w_i^f * x_t + b_i^f + w_h^f * h_{t-1} + b_h^f\right)$$

$$i_t = sigmoid(w_i^i * x_t + b_i^i + w_h^i * h_{t-1} + b_h^i)$$

The decoder also uses an LSTM network for its mapping between the intermediate vector and the output vector. The mathematical representation is same as in encoder. But, here the input to the decoder will be the output from the encoder module. Based on the intermediate vector, it generates the output vector in the target language. The gate values and LSTMCell are calculated using the intermediate vector, which is the output of the encoder.

In case of the sequence to sequence network described above, the contextual information of the word at n^{th} position is passed on to the $(n+1)^{th}$ position. The decoder gets the input from the encoded output of the last sequence of encoder. It is considered that the contextual information of all the preceding sequences is stored in the last sequence of encoder which will help the decoder to generate the sequence of words in target languages. It actually stores the contextual information, but, in case of Indian languages such as Hindi and Tamil, the contextual information of target language has some dependency with any of the sequences in the source language text. Thus, the impact of last words contextual information won’t be sufficient for an accurate translation. Also, there is a need of special alignment phase during this process such that the target text is grammatically correct with respect to the target language grammar. Both these issues have been taken care by introducing the attention mechanism in the sequence to sequence model that too in between the encoder and decoder model. In the attention mechanism [19], there is a method to identify which encoded sequence is important for the decoder during its processing. The attention mechanism for identifying performs a multiplication of the encoded output with the weights so that a weighted vector is formed. The weights are identified using the training of a feed forward network which takes the encoded output as input to it and the decoder output as its output.

5.5. Comparative study of Word sense based Hindi to Tamil machine translation with pivot and without pivot

This word sense based machine translation system has been evaluated using various sentences (restricted to health domain) and the Bilingual Evaluation Understudy (BLEU) score is calculated to be 0.68. Precision and recall gradually increases with respect to corpus size.

Table 2: Comparison of Word sense based Hind-Tamil machine translation with pivot and without pivot

S. No.	Corpus Size (in number of words)	Word Sense Based Hindi-English-Tamil Machine Translation		Word Sense based Hindi-Tamil Machine Translation	
		Precision (in %)	Recall(in %)	Precision (in %)	Recall (in %)
1	10000	68	58	65	63
2	20000	70	69	76	72.5
3	30000	81	76	87	86.5

The figure-6 and figure-7 show the comparison of various systems in terms of precision and recall respectively. It is very clear from the figure-4 that the precision of proposed word sense based Hindi-English-Tamil MT system improves with respect corpus size, but its precision is lesser when compared with word sense based Hindi-Tamil MT system which is due to the increase in distortion by the inclusion of pivot language during translation. From figure-5, it is visible that the recall is relatively good compared to the pivot based Hindi-Tamil MT but degrades when compared to the word sense based Hindi-Tamil MT due to the distortion.

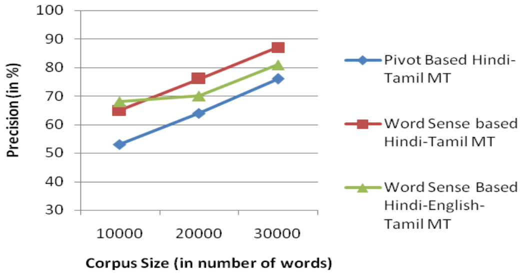

Figure 6: Comparison of Hindi to Tamil Machine Translation in terms of precision

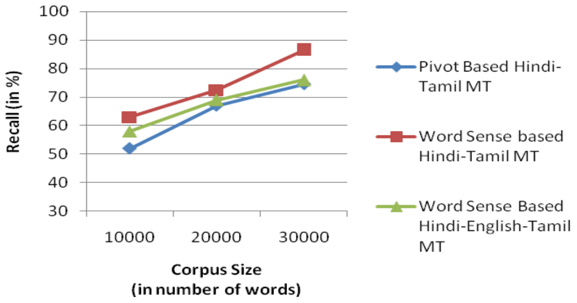

Figure 7: Comparison of Hindi to Tamil Machine Translation in terms of recall

5.6. Comparative study of neural machine translation and statistical machine translation

The neural machine translation system was developed using sequence to sequence learning based and the output of this system has evaluated using the BLEU score. The BLEU score for the various training and testing corpus is as shown in table-III. It is found that the BLEU score increases with respect to the training corpus size. But when the training corpus is 90% of corpus given, it degrades the BLEU score and accuracy reduces. This is since there is a slight deviation in mapping of the source text with the target text due to equal probable chances for multiple target sentences.

Table 3: Result analysis of Neural Machine translation

S. No.	Training/Testing Corpus Size (in %)	BLEU Score
1	60/40	0.7037
2	70/30	0.7234
3	80/20	0.7588
4	90/10	0.6628

6. Organization of Thesis

The thesis has been organized into seven chapters. The brief outline of each chapter is given below:

The chapter 1 discusses the present scenario in machine translation system. It also discusses about the motivation, objective, issues and solution approaches used in the present research. Further, the major issues that has been identified in the existing machine translation system. Finally, the possible solution approach and contribution to the identified problems are discussed.

In chapter 2 focuses on existing related work on machine translation systems. This chapter discusses about various approaches that has been used in a machine translation and its related drawbacks. It also discusses about the existing Indian language processing systems that can aid in translation process.

In chapter 3, the focus is on the word sense based Hindi to Tamil machine translation system. It describes about the phases used in the approach and discusses about the results generated using this approach. This chapter has the description about the short-comings in this approach.

Chapter 4 deals with pivot based Hindi to Tamil machine translation system. It details about the need for a pivot language and how to choose the pivot language. Also describes about the phases used and the result analysis in this approach.

In chapter 5, a hybrid machine translation system was introduced which makes use of both the word sense as well as the pivot language. This chapter describes about the phases and working of the system. The result analysis of this approach was described in this chapter along with its short-comings with respect to the two languages under consideration, Hindi & Tamil

In chapter 6, introduces the deep learning approach for Hindi to Tamil machine translation. It details about the need for such a system and various stages that will be used in this approach. Each stage used in this approach has been described in detail. Results of this approach has been analyzed and compared with the preceding chapter results.

Chapter 7 summarizes the contributions of this research work and indicates the scope for the possible future extensions in this area of research.

7. Conclusion and Future Scope

This research work shows the various sense based machine translation system for the Hindi to Tamil language. In this sense based statistical machine translation system, the accuracy was good when compared with the system without sense disambiguation phase. But when compared with system which does not use a pivot language, the system has degradation in its performance this is due to the increase in noise due to the use of English language as pivot. The noise also gets doubled with the increase in corpus size. Thus, the system performance keeps decreasing if there is increase in corpus size. Ideally, it has been identified that the corpus size can be kept at 30,000 words. To improve the accuracy of the system, a neural machine translation system was proposed, which makes use of sequence to sequence learning. The BLEU score of the neural machine translation system is found to be 0.7588 whereas, the word sense based statistical machine translation using a pivot language is found to be 0.68. There is a remarkable improvement in the accuracy of the system when neural machine translation is used.

There can be a good improvement in the sense based statistical machine translation system's performance if the quality of data is improved with the increase in corpus size. Use of quality data can help in reducing the distortion that is occurring due to the usage of pivot language in between. It can also improve the performance by using a pivot language which is syntactically/semantically related to both the source and target language. For example, in case of Hindi and Tamil language, Sanskrit language can be used as a pivot. But only bottleneck is the resource availability in Sanskrit. This system can also be extended further by using some other approach of word sense disambiguation. The neural machine translation system can be improved further using a suitable deep learning based word to vector model so that the semantics are captured in an efficient manner. Even the accuracy of this system can be improved by improving the quality of corpus that is being used.

Publication from Research

1) Vimal Kumar. K, Divakar Yadav, 2018: Word Sense Based Approach for Hindi to Tamil Machine Translation Using English as Pivot Language. *Journal on Advanced Intelligence Paradigms,* doi: 10.1504/IJAIP.2018.10008778. (h-index [**6**], h5-index **[7]**, h5-median [**8**]), SJR=0.199, ***(serial no. 2498 in UGC list of Journals).***

2) Vimal Kumar. K, Divakar Yadav, 2018: Word Sense Based Hindi-Tamil Statistical Machine Translation. *Journal on Intelligent Information Technologies,* doi: 10.4018/IJIIT.2018010102. (h-index: [10]), SJR=0.239, ***(serial no. 22986 in UGC list of Journals).***

3) Vimal Kumar. K, Divakar Yadav, 2015: An improvised extractive approach to Hindi text summarization, *International conference on Information System Design & Intelligent Applications*, Advances in Intelligent and soft computing (AISC) – Springer, Vol. 339, ISBN: 978-81-322-2249-1, ISSN: 2194-5357, pp. 291-300, 8-9 Jan 2015. doi:10.1007/978-81-322-2250-7_28. **(Cited by 6), Indexed in ISI Proceedings, EL-Compendex, DBLP, SCOPUS, Google Scholar and SpringerLink.**

4) Vimal Kumar. K, Divakar Yadav, Arun Sharma, 2015: A Graph based technique to Hindi text summarization, *International conference on Information System Design & Intelligent Applications*, Advances in Intelligent and soft computing (AISC) – Springer, Vol. 339, ISBN: 978-81-322-2249-1, ISSN: 2194-5357, pp. 301-310, 8-9 Jan 2015. doi:10.1007/978-81-322-2250-7_29. **(Cited by 4), Indexed in ISI Proceedings, EL-Compendex, DBLP, SCOPUS, Google Scholar and SpringerLink.**

5) Vimal Kumar. K, Yamuna Prasad, 2019: A sequential approach to handle machine translation of low resource languages, *In the proceedings of 20th International conference on Computational linguistics & Intelligent text processing (CICLING), pp XX-XX, 2019,* **Indexed in SCOPUS, DBLP, Web of Science.** (h5-index [**21**], h5-median [**33**]).

References

[1] Paul, M., Finch, A., and Sumita, E, How to choose the best pivot language for automatic translation of low-resource languages. ACM Transaction on Asian Language Information Processing, 12, 4, Article 14 (October 2013), 17 pages.

[2] Huang, C.-C., Chen, M.-H., Yang, P.-C., and Chang, J. S., A computer-assisted translation and writing system. ACM Transaction on Asian Language Information Processing, 12, 4, Article 15 (October 2013), 20 pages

[3] Akshar Bharati, Chaitanya Vineet, P. Amba Kulkarni, Rajeev Sangal, Anusaaraka,"Machine translation in Stages", A Quarterly in Artificial Intelligence, Vol. 10, No. 3, NCST, Mumbai, pp. 22-25, (July 1997).

[4] Hemant Darbari,"Computer-assisted translation system – an Indian perspective", Machine Translation Summit VII, 13th-17th September Kent Ridge Digital Labs, Singapore. pp. 80-85, 1999.

[5] R.M.K. Sinha,"An Engineering Perspective of Machine Translation", AnglaBharti-II and AnuBharti-II Architectures. In proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004). November 17-19. Tata McGraw Hill, New Delhi. pp. 134-38, 2004.

[6] R. M. K. Sinha, Anil Thakur,"Machine Translation of bi-lingual Hindi-English (Hinglish) text in proceedings of the tenth Machine Translation Summit", MT Summit X, Phuket, Thailand, September 13-15. pp.149-156, 2005.

[7] Vishal Goyal, Language in India, Ph.D. Thesis on, "Development of a Hindi to Punjabi Machine Translation System" [Online] Available: http://www.languageinindia.com 599 10: 10 October 2010.

[8] Vishal Goyal, Gurpreet Singh Lehal,"Web Based Hindi to Punjabi Machine Translation System", journal of emerging technologies in web intelligence, Vol. 2, May 2010.

[9] Rahul Kumar Yadav and Deepa Gupta (2010) 'Annotation Guidelines for Hindi-English Word Alignment', International Conference on Asian Language Processing.

[10] Raju Korra, Pothula Sujatha, Sidige Chetana, Madarapu Naresh Kumar (2011) 'Performance Evaluation of Multilingual Information Retrieval (MLIR) System over Information Retrieval (IR) System', IEEE-International Conference on Recent Trends in Information Technology (ICRTIT).

[11] Ananthakrishnan, R., Bhattacharyya, P., Hegde, J. J., Shah, R. M., and Sasikumar, M. (2008) 'Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation', Proceedings of International Joint Conference on Natural Language Processing.

[12] S. Lakshmana Pandian and Dr. T. V. Geetha, Morpheme based Language Model for Tamil Part-of-Speech Tagging**,** In Research Journal on Computer Science and Computer Engineering with Applications (POLIBITS08), Issue 38, July-Dec 2008, pp. 19-25

[13] Imam, A.H.; Arman, M.R.M.; Chowdhury, S.H.; Mahmood, K., "Impact of corpus size and quality on English-Bangla statistical Machine Translation system," in

Computer and Information Technology (ICCIT), 2011 14th International Conference on , vol., no., pp.566-571, 22-24 Dec. 2011

[14] Sukhoo, A.; Bhattacharyya, P.; Soobron, M., "Translation between English and Mauritian Creole: A statistical machine translation approach," in IST-Africa Conference Proceedings, 2014 , vol., no., pp.1-10, 7-9 May 2014

[15] Rui Wang; Hai Zhao; Bao-Liang Lu; Utiyama, M.; Sumita, E., "Bilingual Continuous-Space Language Model Growing for Statistical Machine Translation," in Audio, Speech, and Language Processing, IEEE/ACM Transactions on , vol.23, no.7, pp.1209-1220, July 2015

[16] DeyiXiong; Min Zhang; Xing Wang, "Topic-Based Coherence Modeling for Statistical Machine Translation," in Audio, Speech, and Language Processing, IEEE/ACM Transactions on , vol.23, no.3, pp.483-493, March 2015

[17] S. Lata, S. Chandra, Development of Linguistic Resources and Tools for providing multilingual Solutions in Indian Languages - A Report on National Initiative, in Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), 2010.

[18] A. Bharati, S. Husain, D. M. Sharma, and R. Sangal, Two stage constraint based hybrid approach to free word order language dependency parsing. In Proceedings of the 11th International Conference on Parsing Technologies (IWPT09). Paris. 2009.

[19] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473 [cs.CL], September 2014.

[20] C. Feng, T. Li, and D. Chana, “Multi-level anomaly detection in industrial control systems via package signatures and lstm networks,” in Dependable Systems and Networks (DSN),2017-47thAnnualIEEE/IFIP International Conference on. IEEE, 2017, pp. 261–272.

[21] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). MIT Press, Cambridge, MA, USA, 3104-3112.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1 (January 2014), 1929-1958.