Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Advanced Algorithm for Matching & Reordering in EBMT System (English to Hindi Translation)

[1]Pranshu Bhardwaj, [2]Romsha Vishwakarma, [3]Shashi Pal Singh, [4]Ajai Kumar, [5]Hemant Darbari
[1, 2] Banasthali Vidyapith, Banasthali, India
[3, 4, 5] AAI, Center for Development of Advanced Computing, Pune, India

Abstract-- Example Based Machine Translation performs translation after finding and combining sub-sentential matches from the training corpus. It is based on principle: remember everything translated in the past and use everything available to facilitate the translation of the next utterance. Sometime it is tedious task to find out matched chunk and phrase from various sentences which are in exiting corpus, but are not aligned as the source sentences or as in training corpus.
In this paper, we propose an alleviated algorithm or method to handle this problem by creating chunks with syntactic approach of extracting and matching the chunks followed by the reordering algorithm of linguistics structures.

Keywords- Example Based Machine Translation, Bilingual Corpus, Synthesizer, Reordering, N-grams, Matching, Alignment and etc.

I. INTRODUCTION

In today's era, the world is following the way of "Vasudhaiv Kutumbkam" and development of different technologies & Science made it easier for common people over the world wide.
For transmission of thoughts & ideas Language is always a main medium and it is not possible for human being to have the knowledge of every language & their accent as well, used by different people in different regions and countries.
Machine Translation has overcome this Linguistic difficulty by providing Translations of One Language into another by making the Computer Artificially Intelligent.

A. Why EBMT

Example Based Machine Translation is an approach to guide people about translation with the help of Existing Examples as in normal human life, a baby is made intelligent or provided knowledge by practicing some examples (Approach of Learning by Examples).
For common People, who are not related with Science or IT can learn this approach without any hazarders of Rules and Structures.

B. Problems in English to Hindi Translation

First problem is, in different Foreign Languages (out of India), it is found that many of the languages have 90% one to one corresponding with English Language
As in French-English:

French	Mon 1	Prenom 2	est 3	Ram 4
English	My 1	name 2	is 3	Ram 4

French	Puis 1	J'entre' 2	Monsiour 3	S'ilvous plait 4
English	May 1	I come in 2	sir 3	please 4

In German-English:

German	Ich 1	habe 2	ein 3	Stift 4
English	I 1	have 2	a 3	Pen 4

If we translate from French to English or German to English or vice-versa, there is no requirement of reordering and corresponding meaning match in most of the sentences as they are placed at same index in both the languages.

But in Hindi, the most fluctuating language of world, it is a tedious task to reorder the sentence in its correct structure and to make the word so intelligent that it can get its meaning at correct index during translation.

English	I 1	live 2	in 3	India 4
Hindi	मैं 1	भारत 4	में 3	रहताहूँ 4

Second problem is Hindi Sentence generally ends with an Auxiliary Verb but in English Language it is not necessary. As in above example, there is no auxiliary in English but Hindi Sentence has "हूँ".

Third problem in English to Hindi translation is, due to fluctuation Hindi sentences can be written & spoken in different structures, as:

आपकैसेहोयाकैसेहोआप

In the above example, both sentences are correct and they both have same meaning.

Fourth problem, in India itself different regions have variations in their Hindi accent.So it is necessary to have a good translator of English to Hindi.

II. MAKING SYSTEM ARTIFICIALLY INTELLIGENT

The System is made intelligent by rich bilingual corpus of Hindi-English. It includes Bilingual sentences, phrases (Sayings, Phrasal Verbs and PhrasalNouns etc.) and words. Words are assigned categories by using Stanford Tagger and Synonyms in Hindi [9].

III. APPROACH

This System follows EBMT approach for translation. The Input English Sentence is translated with the help of existing translations.

Example- Ram and Shyam are going to CDAC.

Existing Examples in Corpus:

Ram and Shyam are playing.	**रामऔरश्याम**खेलरहे**हैं**।
Anshika goes **to CDAC**.	अंशिका**सीडैकको**जातीहै।
They **are going to** market.	वेबाजार**कोजारहेहैं**।

Result:रामऔरश्यामसीडैककोजारहेहैं।

The System will extract all the words from existing examples which are present in input sentence and will train itself by learning their corresponding meaning in Hindi to provide the translation[1], [2].

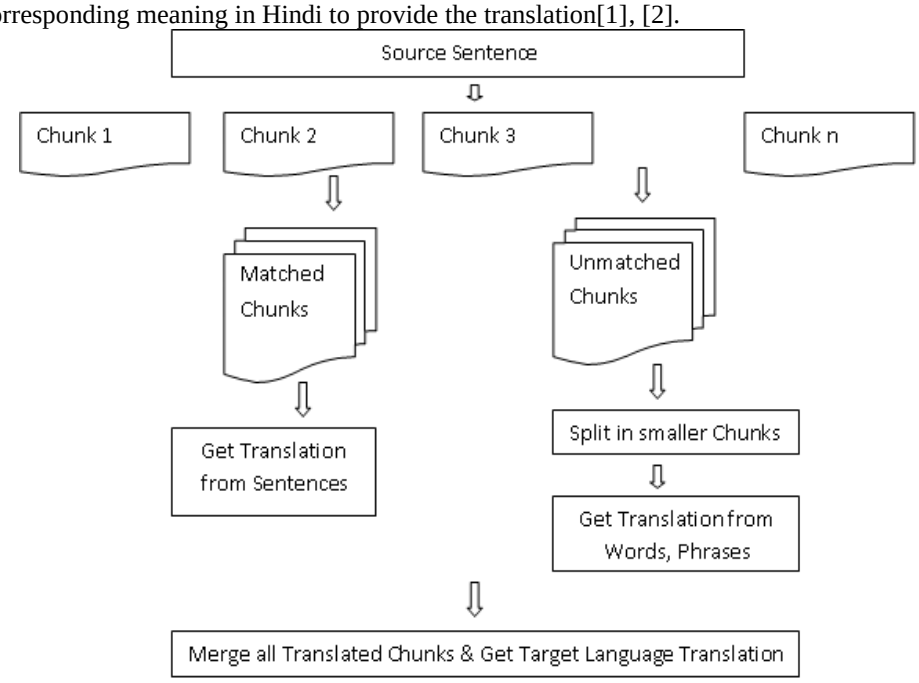

Fig.1. Chunk Based EBMT Approach

In this paper, we are mainly focused on Chunk based EBMT System. The sentence is divided in chunks (with the help of N-grams) and they are get translated [5], [6].

Begin:
Step 1- Take an input English sentence.
Step2- Preprocessing /* (Input Sentence is refined, words are POS Tagged) */
Step 3- (3.1) If input sentence is exactly present in Bilingual Sentence Corpus then it is directly translated into Hindi according to corpus. else: (3.1) N-gram matching is performed. (3.2) Alignment is done according to indices. (3.3) Reordering of aligned Hindi sentence is performed according to the Target Language Structures. (3.4) synthesize the reordered Hindi Sentence with the help of Noun-Adjectives. (3.5)Synonym Handling is performed for semantically correct results.
Step 4- Translated Hindi Sentence is printed as result.
End:

Fig.2. Chunk Based EBMT Algorithm

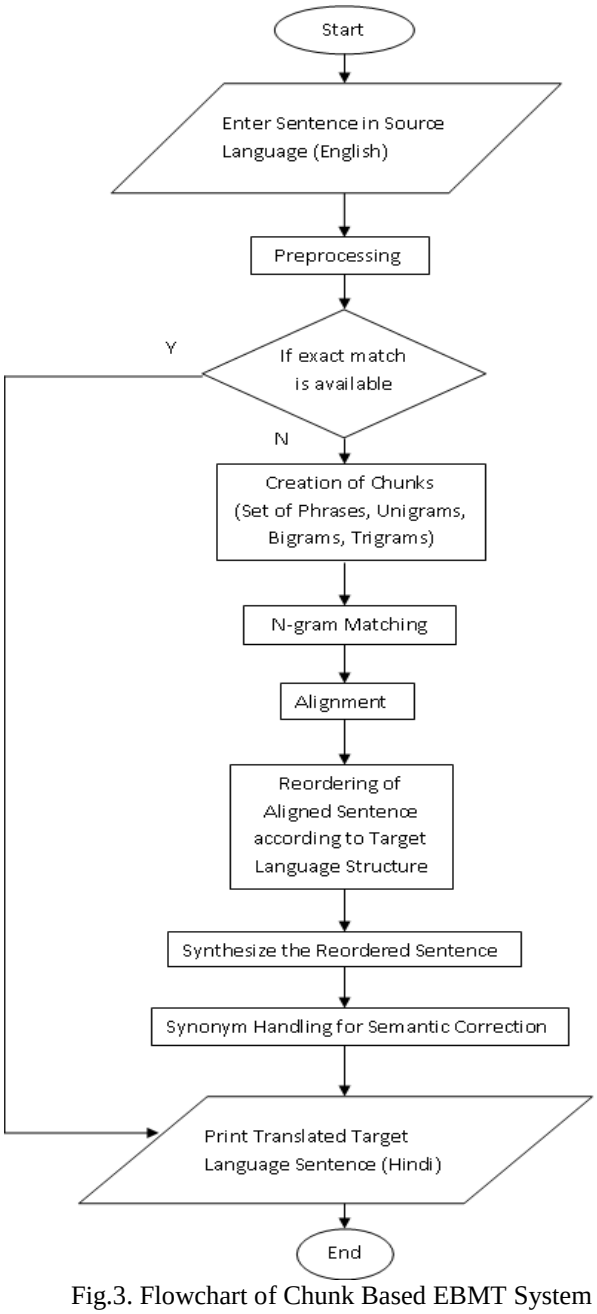

Fig.3. Flowchart of Chunk Based EBMT System

V. ARCHITECTURE OF THE SYSTEM

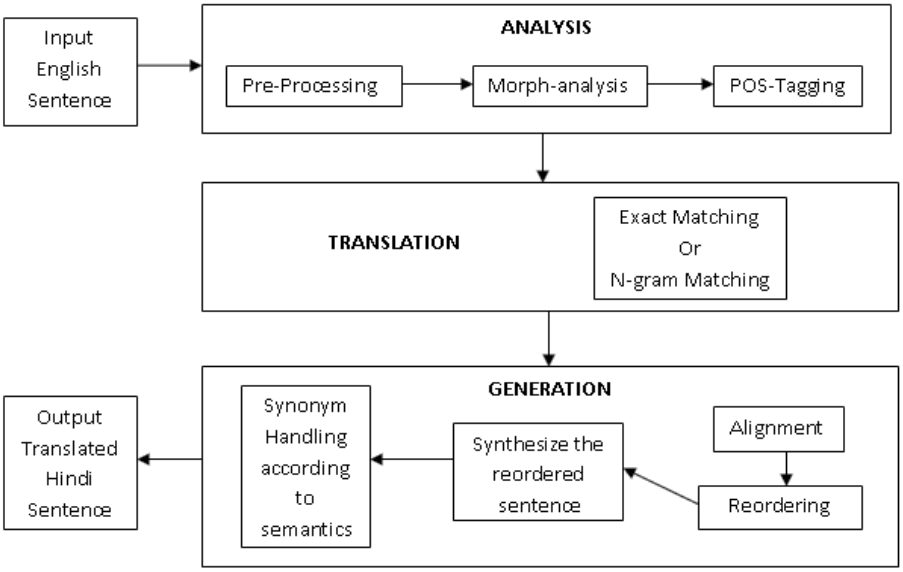

Fig.4. Architecture of Chunk Based EBMT System

A. Pre-Processing:

In its purest form, there is no pre-processing of the corpus in EBMT: everything is done at run time.
Rich bilingual corpus is built for sentences, phrases & words.

B. POS Tagging:

All the words of corpus are tagged with the help of Stanford Tagger. Words are categorised as NOUN, VERB, ADJ, NUMBER, ADV, DET, AUX etc. If word is not found it is transliterated and then it isPOS-Tagged on run time with the help of Stanford Tagger.
In English sentences, some clauses or phrases (Verbal Phrase, Idioms and Sayings etc) are POS Tagged as PH. They are stored in database separately, with their Real world meanings, as

cats and dogs	मूसलाधार
Good for nothing	निकम्मा

He is going to school	He<PRON> is<AUX> going<VERB> to<PREP> school<NOUN>
Mohan is my best friend	Mohan<NOUN> is<AUX> my<PRON> best<ADJ>friend<NOUN>
it is raining cats and dogs	It<PRON> is<AUX> raining<NOUN>cats and dogs<PH>
I have blue eyes and fair hair	I<PRON> have<AUX> blue<ADJ> eyes<NOUN> and<CONJ> fair<ADJ> hair<NOUN>

C. Matching:

In this process all the Source sentences of corpus are indexed with their corresponding Target Sentences with the help of their meanings.
Every English Sentence word is provided the index of its existing word meaning in Target Sentence. Thus the Index Column for each English Sentence contains the string of indices of its words' meaning index in Target Sentence [3].
Algorithm:

Begin:
Step 1- Take the preprocessed input English Sentence.
Step 2- Create Bigrams of the input sentence.
Step 3- Scan the corpus sentences for matching the bigrams.
Step 4- Fetch the corresponding bigram Hindi meaning from the bilingual corpus.
Step 5- If bigrams are not matched then split the non-matched bigrams in unigrams and repeat Step 3 & Step 4.
Step 6- If Unigrams are not matched in sentence corpus then find the Hindi meaning of unigrams from bilingual word corpus.

Step 7-[Alignment] Put the Hindi meaning of bigrams, unigrams & words according to the English word indices. (Translated Hindi Sentence)
End:

Fig.5. Algorithm for Index Matching and Alignment

Ram and Shyam are playing. 0 1 2 3 4	रामऔरश्यामखेलरहेहैं। 0 1 2 3 4	0 1 2 4 3
Anshika goes to CDAC. 0 1 2 3	अंशिकासीडैककोजातीहै। 0 1 2 3	0 3 2 1
She is singing. 0 1 2	वहगारहीहै। 0 1 2	0 2 1

Input:
Ram and Shyam are going to CDAC.
Creation of Bigrams:

Ram and	and Shyam	Shyam are	are going	going to	to CDAC.

Matched Sentences:

Ram and Shyam are playing.	**रामऔर**श्यामखेलरहेहैं।	**0 1** 2 4 3
Ram **and Shyam** are playing.	राम**औरश्याम**खेलरहेहैं।	0 **1 2** 4 3
Ram and **Shyam are** playing.	रामऔर**श्याम**खेलरहे**हैं**।	0 1 **2 4** 3
They **are going** to market.	वेबाजारको**जारहेहैं**।	0 **4 3** 2 1
They are **going to** market.	वेबाजार**कोजारहे**हैं।	0 4 **3 2** 1
Anshika goes **to CDAC.**	अंशिका**सीडैककको**जातीहै।	0 3 **2 1**

D. Alignment:

Ram and Shyam are going to CDAC. रामऔरश्यामहैंजारहेकोसीडैक
0 1 2 3 4 5 6 0 1 2 3 4 5 6
If any bigram is not matched in existing examples then it is split in next lower level N-gram as bigram (B) to unigram (U).
input:
Ram and Shyam are going to garden.
Creation of Bigrams:

Ram and	and Shyam	Shyam are	are going	going to	to garden.

Matched Sentences:

Ram and Shyam are playing.	**रामऔर**श्यामखेलरहेहैं।(B)	**0 1** 2 4 3
Ram **and Shyam** are playing.	राम**औरश्याम**खेलरहेहैं। (B)	0 **1 2** 4 3
Ram and **Shyam are** playing.	रामऔर**श्याम**खेलरहे**हैं**। (B)	0 1 **2 4** 3
They **are going** to market.	वेबाजारको**जारहेहैं**। (B)	0 **4 3** 2 1
They are **going to** market.	वेबाजार**कोजारहे**हैं। (B)	0 4 **3 2** 1
They are going **to** market.	वेबाजार**को**जारहेहैं।(U)	0 4 3 **2** 1
They play in **garden.**	वे**बगीचे**मेखेलतेहैं।(U)	0 3 2 **1**

Alignment:

Ram and Shyam are going to garden. 0 1 2 3 4 5 6	रामऔरश्यामहैंजारहेकोबगीचे 0 1 2 3 4 5 6

If any unigram is not matched in existing examples then that unigram is given meaning from word bilingual corpus.
Ram and Shyam are going to garden **for enjoy**.

for	केलिये
enjoy	आनन्द

Ram and Shyam are going to garden **for enjoy**. 0 1 2 3 4 5 6 7 8	रामऔरश्यामहैंजारहेकोबगीचे**केलियेआनन्द** 0 1 2 3 4 5 6 7 8

E. Reordering:

After performing Matching of the input sentences according to the existing corpus, linguistic rules for proper Hindi Translations, are applied on aligned chunks for reordering.

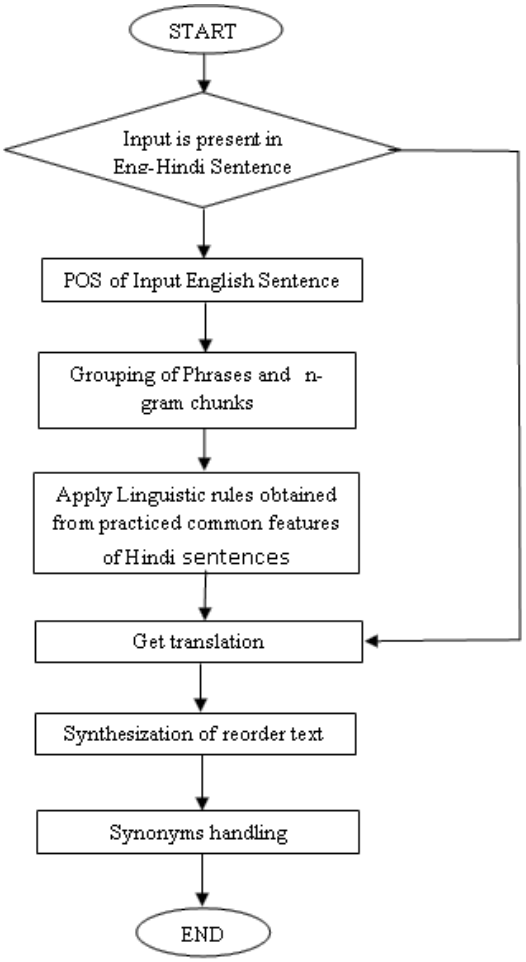

Fig.6. Flowchart for Reordering

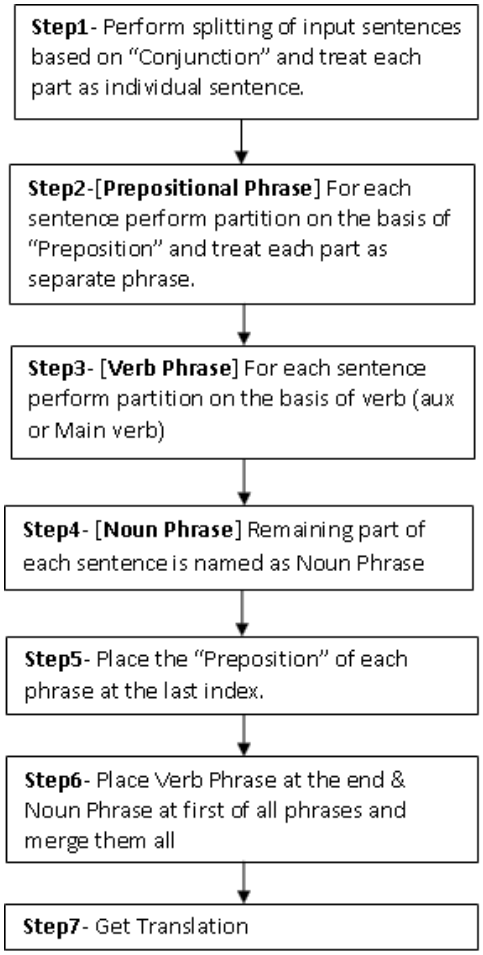

Fig.7. Linguistic Rules for Reordering

The basic idea behind this method is to divide the input sentence by "Conjunction" then by "Preposition" and finally by "Verbs" (AUX or Main Verb)[10],[11].

Input- Ram is going to school.

POS tagging	NOUN1 AUX VERB PREP NOUN2
Reordering (according to Hindi Linguistic Rules)	NOUN1 NOUN2 PREP VERB AUX
Output	रामविद्यालयकोजारहाहै

Begin:
Step1: Take English Sentence as input
Step2: if Input Sentence already exists, directly give its Hindi translation as output and exit. else Go to step3
Step3: if CONJ==true Break the sentence into separate sentence else go to step4
Step4: if PREP==true Split the sentence into Prepositional Phrases else go to step5
Step5: if (in remaining part after splitting the Prepositional Phrase) AUX or VERB==true Split the remaining Sentence in Noun Phrase and Verb Phrase else if AUX or VERB==false Remaining part of sentence is Noun Phrase else go to step6
Step6: Place PREP of each Prepositional Phrase at the last index of that phrase.
Step7: Place AUX of each Verb Phrase at the last index of that phrase & VERB at the second last index of that phrase.(If AUX is not present place VERB at the last index)
Step8: Keeping Noun Phrase at first then Prepositional Phrases and at last Verb Phrase Merge all the Phrases.
Step9: Get the Translation of Each corresponding English word in Hindi with the help of POS Tags.
Step10: Print the Translation.
End:

Fig.8. Algorithm for Reordering of Aligned Sentence

Example: He is trying to help.
PRON AUX VERB PREP VERB
NP VP PP
Step3: Since no CONJ goto step2.
Step4:
P1: PRON AUX VERB
P2: **PREP**VERB
Step5:
P1: PRON
P2: **AUX**VERB
P3: **PREP** VERB
Step6:
P1: PRON //NP
P2: **AUX** VERB //VP
P3: VERB **PREP** //PP
Step7:
P1: PRON
P2: VERB **AUX**

P3: VERB **PREP**
Step8:
PRON VERB PREP VERB AUX
NP PP VP
Step9: He<PRON> is<AUX> trying <VERB> to<PREP> help<VERB>
PRON VERB PREP VERB AUX
He help to trying is
NP PP VP
वहमददकोकोशिशकररहाहै

In cases where more than one prepositions exist, place them in order given below-
Example: He is going to new school for better education.
P1- he
P2- is going
P2- to new school
P3-for better education
And then place the later preposition phrase before earlier one and then follow the same procedure.
P1-he
P2-is going
P2-for better education
P3-to new school

There are sentences in languages which contain various words in a Phrase (Prepositional, Verb etc) [4], [12], then it is not possible to get a proper Translation by just having PREP at the last index of Prepositional Phrase & AUX -VERB at last index of Verb Phrase. Sometimes many English intermediate words are needed to change their indices.
So for resolving this problem we can make intelligent our system by providing the Phrasal Clauses for Hindi as well.

English Phrases	Hindi Phrases
NOUN Nom1 PREP VERB AUX VERB DET NOUN Nom1 PREP **VERB PRON** AUX **PRON NOUN ADV1 ADV2 ADV3**	NOUN Nom1 VERB PREP DET NOUN Nom1 VERB AUX **PRON VERB** PREP **PRON ADV3 ADV2 ADV1 NOUN** AUX

He	learns	from books very easily
PRON	VERB	PREP NOUN ADV1 ADV2
NP	VP	PP

In above Example, Prepositional Phrase for Target language cannot be getting by just putting PREP at last index.

PREP NOUN ADV1 ADV2	NOUN ADV1 ADV2 PREP	NOUN PREP ADV1 ADV2
	wrong	**right**

Target Language Structure:

NP	PP	VP
PRON	NOUN PREP ADV1 ADV2	VERB
वह	किताबोसेबहुतजल्दी	सीखताहै

VI. EVALUATION & TESTING

The N-gram based approach can be more beneficial in Example Based Machine Translation. We can have more accurate translations by matching chunks from existing examples of bilingual corpus. If we align these chunks then we can have high quality results.
For this evaluation we used 11221 bilingual sentences' corpus and 13500 bilingual words' corpus and about 1500 Phrases (Verbal Phrase, idioms & Sayings).
For testing we took 1000 sentences in English Language and performed Example Based Machine Translation using N-grams, Matching Index technique and Reordering by Linguistic Rules of Hindi Language.

VII. RESULTS & ANALYSIS

Sentences are 75-80% accurate using above Chunk based EBMT. About 450 sentences are 80-100% accurate, about 440 sentences are 60-80% accurate & remaining sentences are below 60% accurate.

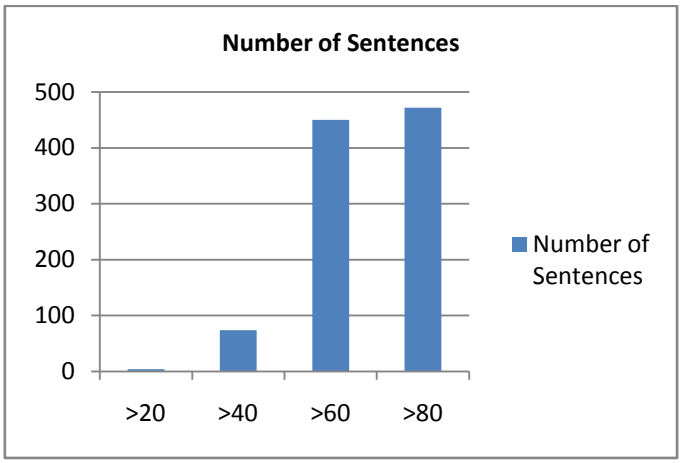

Fig.9. Graphical representation of accuracy of sentences in %

Although, Chunk Based Machine Translation, Index Matching has given results in its Positive contributions but there are some points which need to be enhanced further. In some translations, Case Markers are missing, which are generally used in Hindi Sentences for providing the basic structure and support in joining the Hindi words for their appropriate Translations.

"Ram ate mango" will give Translation "रामआमखाया" but "रामनेआमखाया" is expected for appropriate translation and in another case, English Sentences have Prepositions which are not generally used in Hindi Language but are grammatically correct.

"He went to abroad for studies" will give Translation "वहपढ़नेकेलिएविदेशकोगया"but in generally it is taken as "वहपढ़नेकेलिएविदेशगया". These are some point where this system deviates from its accuracy.

VIII. FUTURE WORK

In this paper, we have given some strategies for enhancing and making easier the task of Matching and reordering for English to Hindi EBMT System. From Chunk analysis, we recognize the consistent chunk sequences in the corpus and their corresponding Hindi Translation easily.

For Case Markers Handling,English words which can be used as VERB and NOUN both etc are needed to take in consideration for Translation accuracy. We used relatively small data for experiment and we think that some more improvements can be taken in the method for making the translation more appropriate.

REFERENCES

[1] Chunyu Kit, Haihua Pan and Jonathan J. Webster, "Example-Based Machine Translation: A New Paradigm"

[2] Ruchika A. Sinhal & Kapil O. Gupta, "A Pure EBMT Approach for English to Hindi Sentence Translation System"-*I.J. Modern Education and Computer Science*, 2014, 7, 1-8 (2014)

[3] Khan Md. Anwarus Salam, Setsuo Yamada, Tetsuro Nishino, "Example-Based Machine Translation for Low-Resource Language Using Chunk-String Templates"- 2011

[4] Media A. Ayu, Teddy Mantoro, "An Example-Based Machine Translation Approachfor Bahasa Indonesia to English:An Experiment Using MOSES"-*2011 IEEE Symposium on Industrial Electronics and Applications (ISIEA2011)*, September 25-28, 2011, Langkawi, Malaysia

[5] Jae Dong Kim, "Chunk alignment for Corpus-Based Machine Translation"-*CMU-LTI-11-002*September 29, 2010

[6] Jae Dong Kim Ralf D. Brown Jaime G. Carbonell, "Chunk-Based EBMT"-*EAMT May 2010 St Raphael*, France

[7] R. Mahesh K. Sinha, "Developing English-Urdu Machine Translation Via Hindi"- 2009

[8] HAROLD SOMERS "Review Article: Example-based Machine Translation"- 1999

[9] Nisheeth Joshi and ItiMathur, "Design of English-Hindi Translation Memory for Efficient Translation"

[10] Vimal Mishra and R. B. Mishra, "Study of Example Based English to Sanskrit Machine Translation"

[11] Rashmi Gangadharaiah, N. Balakrishnan, "Application of Linguistic Rules to Generalized Example Based Machine Translation for Indian Languages"

[12] Jurafsky and Martin, *Speech and Language Processing-An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice-Hall, Inc.2000

