An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism

J. Sangeetha[#1], S. Jothilakshmi[*2] ,R.N.Devendra Kumar [*3]
[1,2]Department of CSE, Annamalai University,
Annamalai Nagar,Chidambaram-608002,
Tamilnadu, India.
[3]Department of CSE,Sri Ramakrishna Institute of Technology,
Coimbatore-641010.
[1]sangita.sudhakar@gmail.com
[2]jothi.sekar@gmail.com
[3]devendrakumar.cse@srit.org

Abstract–Machine Translation is an essential approach for localization, and is especially appropriate in a linguistically diversenation like India. Automatic translation between languages which are morphologically rich and syntactically different is generally regarded as a complex task. A number of machine translation systems have been proposed in literature. But, conventional rule-based machine translation system is costly in terms of formulating rules. It introduces inconsistencies, and it is inflexible to be robust. Statistical MT is an approach that automatically attains knowledge from a vast amount of training data. This approach is characterized by the use of machine learning techniques. But, still there is scope for better performance of the system. In this paper, a Hybrid Machine Translation (HMT) approach is proposed which is the combination of rule based and statistical technique for translating text from English to Indian languages such as Tamil, Malayalam and Hindi. The rule based machine translation technique, involves the formation of rules which helps to re-order the syntactic structures of the source language sentence along with its dependency information which brings close to the structure of the target sentence. The parser identifies the syntactical elements in English sentences and suggests its Indian languages translation taking into account various grammatical forms of those Indian languages. Context Free Grammars (CFG) is used in generation of the language structures, and then the errors in the translated sentences are corrected by applying a statistical technique. Simplifying and segmenting an input language text becomes mandatory in order to improve the machine translation quality. The experimental results show that the proposed approach competes with the machine translation methods reported in the literature and it provides the best translated output in each language.

Keyword-Machine Translation (MT)1, Natural Language Processing2, Rule Based Machine Translation (RBMT)3, Statistical Machine Translation (SMT)4,Parsing5.

I. INTRODUCTION

Machine Translation (MT) mainly deals with the transformation from one natural language to another. Natural Language Interface provides the user freedom to interact with the computer in a natural language like English, Malayalam, Telugu, and Hindi or any other language used for day to day communication. One of the important goals of computational linguistics is a fully automatic machine translation between such natural languages. This is important because communication between people from different linguistic backgrounds still poses as a major problem (S. Samantaray. 2007). In earlier days the machine translation is done only at the word level i.e. word by word translation. The three major techniques involved in machine translations are rule based, statistical, and example based techniques. The statistical and example based techniques need parallel corpora for translation. But for Indian languages such as Tamil, Malayalam and Hindi has only few parallel corpora available. In such cases adopting these techniques will not result in proper translation in the target language.

Natural language processing, and artificial intelligence (AI) in general, have focused mainly on building rule-based systems with carefully handcrafted rules and domain knowledge. For a rule-based approach, knowledge is endorsed by linguistic experts and is encoded in termsof rules, and knowledge is represented deeper, and quality and fluency of the translation are better. But a huge amount of fine-grained knowledge is usually required to translate well; it is quite difficult for a rule-based approach to acquire such kinds of knowledge. In addition, the maintenance of consistency among the inductive rules is by no means easy, and the coverage is lower. Therefore, a rule-based approach, in general, fails to attain satisfactory performance for large-scale application.

In contrast, a statistic-based approach provides an objective measuring function to evaluate all possible alternative structures in terms of a set of parameters. Generally, parameters are estimated from a training corpus by using well-developed statistical theorems. The linguistic uncertainty problems can thus be resolved by a solid mathematical basis. Moreover, the knowledge acquired by a statistic-based method is always consistent because all the data in the corpus are jointly considered during the acquisition process. But knowledge is represented shallower than the rule based approach, and quality and fluency are worse than the rule based method. From the above, a rule-based approach and a statistic based approach all have advantages and disadvantages, so the two approaches are combined. In a machine translation system, where the underlying grammar is large, there are many sources which may cause the system to become highly ambiguous.

In this paper we propose a hybrid technique (combination of rule based technique and the statistical technique) to translate the text from English to Tamil, Malayalam and Hindi. The various challenges in this work are:

1. Indian languages are morphologically quite rich and mostly follow the SOV order but the order is flexible according to the sentences.
2. It is difficult to construct the meaningful sentences which are grammatically correct with one word, for example in Tamil "odinaan" shows the tense, action performed, and the gender of the person performing that action. So implementing the gender ending is a difficult task in some Indian languages.
3. Ambiguity problem (H. Yang et al. 2011) arise i.e. words that have more than one meaning when they occur in parts of speech cannot be translated accurately.

Barriers in achieving good quality MT can be categorized into two:

1. Problems that arise due to inherent ambiguities in natural languages and
2. Problems that arise due to structural and lexical difference between natural languages.

When a sentence is uttered, *'intonation'* along with world knowledge helps human brain to interpret the meaning of the sentence unambiguously. But this is not the case when it is being in written form. Before analyzing examples containing ambiguities, we categorize the type of ambiguities as structural or lexical. A sentence is said to be structurally ambiguous, if it can have more than one syntactic structure at surface level. In deeper level it can have only one syntactic structure.

The main contribution of this paper concerns the use of hybrid approach (combination of rule based and statistical) to perform the translation from English to Indian languages such as Tamil, Malayalam and Hindi. New reordering rules have been added to the proposed system in order to make the translation more efficient. The performance has been improved thereby incorporating certain features such as simplifying & segmenting the input text, adding new reordering rules and adding syntactic and morphological information to the plain corpus. Since all Indian languages relatively rich in terms of morphology and follows the SOV order, the hybrid mechanism that we have implemented should be applicable to English to all Indian languages translation in general. Thus the proposed hybrid approach offers the best translation results for the English to Indian language machine translation.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 presents the structure of the Indian languages; Section 4 describes the system overview of the proposed Hybrid machine translation system; Section 5 discusses the performance measure of the proposed system and the experimental results. Finally Section 6 gives the conclusions and describes the future work.

II. RELATED WORK

Continuous advances in hardware technology made the researchers of MT to rephrase their conventional approaches to MT. The result was the revival of 'empiricist' or statistical based methods abandoned in the first decade of MT research. Translation probabilities will be used along with 'language model' to decode the target sequence or sentence given in the source. Refinements for initial word-based alignment models, such as phrase-based translation models (Koehn et al. 2003) were proposed. Systems such as EGYPT and Moses systems (Koehnet al. 2007) are the most popular open-source Statistical Machine Translation (SMT) toolkits available in the public domain.

Earlier models of SMT entirely relied on parallel corpus; no other form of linguistic knowledge had been given to those models. Whereas for morphologically rich and resource poor languages, performance of SMT was poor compared to other language pairs that had very simple morphological inflections. Factored Translation Models (Koehn et al. 2007) are trying to address these problems by taking into account linguistic features such as Part of Speech (POS) and morphology. Research also focuses on incorporating syntax into SMT models with the aim of capturing syntactic structures on both sides. Major companies doing corpus-based research includes Google, Microsoft Research and IBM.

Researches have been done in the case of English to Indian languages translation are mostly concentrated on bilingual machine translation systems. English to Malayalam translation system (RemyaRajan

et al. 2009) was based on rule based technique. A novel framework for English to Tamil translation system was proposed by (Harshawardhan et al. 2011). This was a phrase based translation system using translation memory and concept labeling. (Rahul et al. 2009) proposed a system for translating from English to Malayalam which gives better performance than a phrase based system by avoiding parsing of the target language.

Google Translate is a free translation service that offers immediate translations between 65 different languages. It can translate words, sentences and web pages between any mixtures of the supported languages. With Google Translate, information could be made universally available and helpful, in spite of its written language. Fundamentally, Google Translate utilizes a large database of documents that have been already translated into several languages. It scans these documents looking for linguistic patterns; a process called "statistical machine translation" and utilizes these patterns and rules to produce translations (Maritz J.S. 1981). Basically, the machine is constructed through translations that humans have spent a lot of time translating.

One of the key reasons why Google Translate, or any other most of the machine translation tools including Google Translate is not able to ensure the accurate translation is that it cannot find the context within which words are used and also it cannot infer. Take the following sentences in Google translator as an example in fig.1.

English	"Ravi has gone".
Tamil	"Ravi poyirukiraal".
Hindi	"Ravi chalgaya"

Fig.1 Example for Google Translator

Based on the example, Google Translate does not focus on grammatical rules, as its algorithms mostly depend on statistical examination instead of the conventional rule-based study and as such it fails to translate correctly in all the time. But, the original creator, (Franz Josef Och. 2002), has condemned the significance of rule-based algorithms and stated that statistical approaches perform better.

Apertium is a another free/open-source machine translation platform, which was developed to focus on related language pairs however recently it concentrates with more different language pairs (such as English-Catalan). Apertium utilizes a shallow-transfer machine translation engine which processes the input text in stages, as in an assembly line. But this tool does not include the English to Indian languages translation.

III. STRUCTURE OF INDIAN LANGUAGES

Major three Indian languages such as Tamil, Malayalam and Hindi hold almost 40, 35 and 70 million speakers respectively. These languages are holding a unique identity unrelated script and ancient documented histories. Negative and an affirmative voice pertain with verbs. Male female voice is differentiated by ranking priority not by sex, with one priority ranking status indicates superior statue class the other beings of lowest status. Case and number are illustrated by eliminating the nouns. Indian languages are immensely useful in creating the suffixes with nouns and verbs. Subsequently all these three Indian languages are holding their unique alphabets, associated with the Indian alphabet used for Sanskrit. In spite of Malayalam and Hindi are languages of precious in the historical survey, they are resource poor when analyzed by the computational linguistics. In this paper, maximum illuminations depend on Tamil, Malayalam and Hindi languages.

Tamil language is enormous glue centered language with three gender forms namely male, female and neutral. Singular and plural are the two number forms that popularly illuminate inflection based on the gender, number and tense on the commodity of reference among other features. Indian languages are 'Left branching language', in which at the end of the sentences are holding verbs and have posted positions alternated by prepositions. So adjectives, genitive and associated clauses precede their head nouns in a sentence. 'FEELING' is one of the constructive factors and is related to statements of fact versus contingency, supposition, etc. There are four independent feelings that are exposed are: infinitive, imperative, affirmative and negative. Also these two languages have some extra 'modal' forms such as: indicative, conditional, optative, potential, monitory and conjunctive.

The noun phrase (NP) of these languages is not so tedious and has adjectives derived from nouns or verbs and nouns of various sorts that select case endings and post positions. In some cases NP may consist of pronouns, numerals, color terms, deictic particles. NP may contain nominal head or pronoun and may be subsequently by modifiers. Syntactically noun phrases are recognized by their potential to act as subjects, direct objects, indirect objects and compliment of postpositional phrases. Word order plays a vital role in positional languages like English and normally follow right-branching with Subject-Verb-Object sequence. Unlike the English language Tamil, Malayalam and Hindi are syntax of relatively free word sequence language (K.

Narayana Murthy, 2000). These languages are verb final language and all the noun phrases in the sentence normally appear to the left of the verb. The subject noun phrase may also appear in many different positions relative to other noun phrases in the sentence.

The highly agglutinative languages like Tamil and Malayalam nouns and verbs get infected. Many times we need to depend on syntactic function or context to decide upon whether the particular word is a noun or adjective or adverb or post position (Antony P J et al. 2010). This leads to the complexity in bilingual machine translation. A noun may be categorized as common, proper or compound. Similarly, the verb may be finite, infinite, gerund or contingent. The contingent is a special form of verb found only in Kannada and not found in other Dravidian languages. Other parts of speech were also divided into their own subcategories. Parts of speech ambiguity are another important issue that has to be carefully analyzed while designing a machine translation system.

Table 1. PNG - Suffixes in Tamil

Person	Noun	Gender	Past	Present	Future
First	Singular	M/F	En	en	en
	Plural	M/F	Om	om	om
Second	Singular	M/F	Ar	ar	ar
	Plural	M/F	yargal	yargal	yargal
Third	Singular	M/F	Adu	adu	adu
	Plural	M/F	Argal	argal	argal

Table 2. PNG - Suffixes in Hindi

Person	Noun	Gender	Past	Present	Future
First	Plural	M/F	Dha/dhee	hum	yega
	Singular	M/F	“	hey	“
Second	Singular	M/F	“	ho	“
	Plural	M/F	“	hey	“
Third	Singular	M/F	“	hey	“
	Plural	M/F	“	hey	“

The PNG (Person Noun Gender) and the tense marker concatenated to the verb stems are the two important aspects of verb morphology in Indian languages Tamil and Malayalam. The verbal inflectional morphemes attach to the verbs providing information about the syntactic aspects like number, person, case-ending relation and tense. PNG markers play an important role in word formation in South Dravidian languages except Malayalam. The PNG features of the head noun of the subject NP determine the agreement marker of the verb. Usually the South Dravidian language's verbs follow the regular pattern of suffixation. Table 1 & 2 shows the various PNG suffixes for Tamil and Hindi verb root word.

IV. THE PROPOSED SYSTEM

Rule based machine translation for a language is done by developing the rule structures for every possible sentence in the language. The grammar and dictionary creation are the main tasks in the rule based machine translation. The grammar rules and dictionaries once created can be used and never need changing. In order to improve the machine translation quality, splitting and simplifying an input becomes mandatory. The overview of the proposed machine translation system is shown in fig.2.

A. Sentence splitting and simplification

Sentence splitting is the process of segmenting the paragraph and complex sentences into simpler sentences. Many approaches are available for simplifying and segmenting task. Here we are using a rule based technique (Poornima C et al, 2011) to simplify the paragraph and complex sentences. It is based on coordinating conjunction (for, and, not, but, or, yet and so), subordinating conjunction (after, although, because, before, if, since, that, though, unless, where, wherever, when, whenever, whereas, while, why) and connectives like relative pronouns (who, which, whose, whom). Sentence segmentation is expressed as the list of sub-sentences that are portions of the original input sentence. The meaning of the simplified sentence remains unchanged. Characters such as (. , “?” “!”) are used as delimiters. One of the important prerequisites is the presence of delimiters in the given sentence. Initial paragraph splitting is based on delimiters such as “.” and “?”, then the individual sentences are segmented based on connectives, coordinating and subordinating conjunction. This method is useful as a preprocessing tool for improving the quality of machine translation.

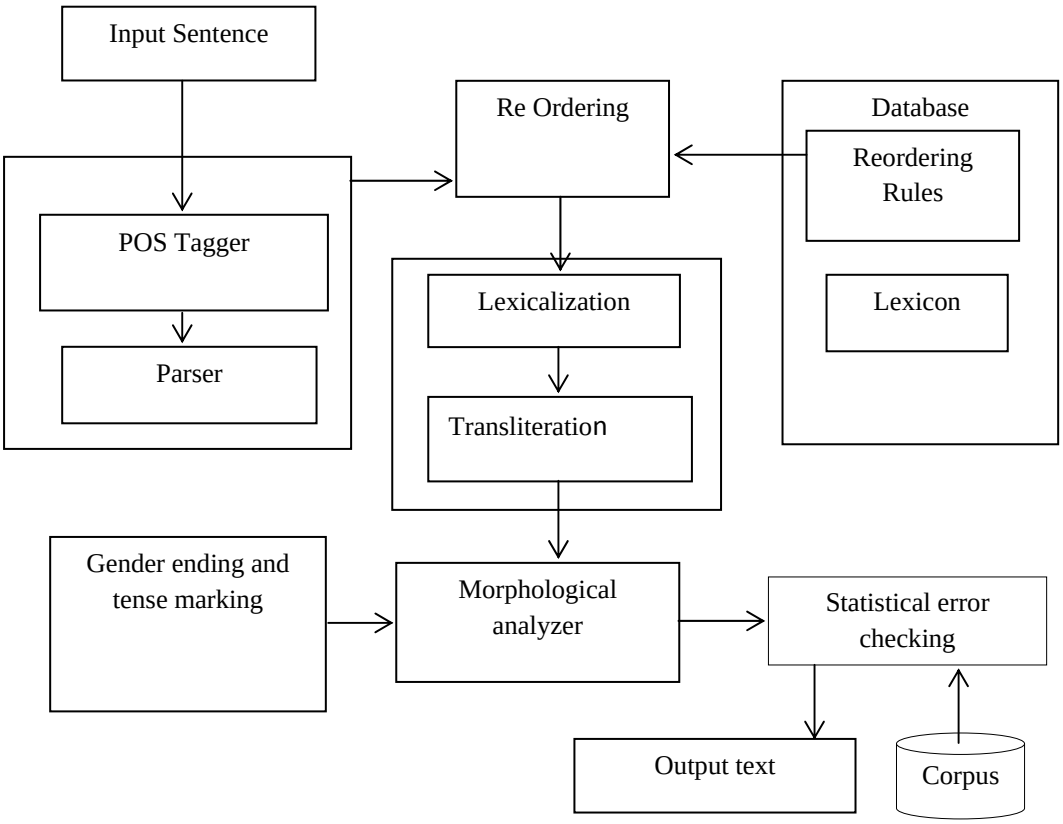

Fig .2 Overview of the proposed system

B. POS tagger

Parts Of Speech (POS) is more significant in case of words because it provides the details about the word's pronunciation and the words morphological affixes. Parts Of Speech tagger is used to tag the source sentences. Tag sets are typically language autonomous. As a result one complete tag set for all languages can be formulated. Typical tag sets will have 20-30 Tags and some have even more. Table 3 presents the different POS tag sets. In this work, we use Stanford POS Tagger for the tagging purpose. The English sentences are fed into the parts of speech tagger. The tagger tokens each word in a sentence and identifies the parts of speech information such as verb, noun, adjective etc. of that word. Then the words and their tagged information are stored in a separate file which is used for reordering according to the Indian language's structure. The output of the POS Tagger for a sentence is given below.

English: They went to the park.

POS *Tag: They/PRP went/VBD to/to the/DT park/NN.*

Table 3. POS Tag set

S.No.	Abbreviation	Type
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, Singular or mass

C. Parsing

Parsing is an important phase which is used to understand the syntax and semantics of any source language sentences confined to the grammar. Parsing is actually related to the automatic analysis of texts according to any grammar. This is done with the help of certain grammar rules that are written as a set of production rules where each production rule is written as:

<Non-terminal>→<Non-terminal | 'terminal' >

A general grammar that can parse an English sentence is illustrated in fig. 3.

<S>	->	<NP><VP>
<NP>	->	<Det><NN>
<VP>	->	<VB><NP><VB>
<Det>	->	'a' | 'an' | 'the'
<NN>	->	'boy' | 'man' | 'apple'
<VB>	->	'ate' | 'run'

Notations:
<S>- Sentence; <NP> -Noun Phrase; <VP>-Verb Phrase; <Det>-Determiner
<NN>-noun; <PP>-Prepositional Phrase
'a', 'an', 'the', 'man', 'apple' –terminals corresponding to the non terminals.

Fig .3 Structure of grammar

In this work, the parse structure of the English sentence is derived using the Stanford parser. The tagged information is passed through the Stanford parser and the parse structure of the sentence is obtained. From this structure we can get the grammar rules for the sentence. Consider the following sentence as an example, "They went to the park". The syntactic parse tree structure for the sentence can be drawn as shown in the fig. 4.

D. Rule based reordering

In machine translation word reordering is a preprocessing step and it makes the translation process easier. Indian languages are the free word-order language. Reordering is the ultimate need for English to any Indian language machine translations. The major structural difference between English and Indian languages is that English follows the structure as Subject-Verb-Object (SVO), whereas, Indian languages follow the default sentence structure as Subject-Object-Verb (SOV). Based on this, languages could be classified as SVO (English), SOV (Tamil, Malayalam, and Hindi), VSO (Arabic), etc.

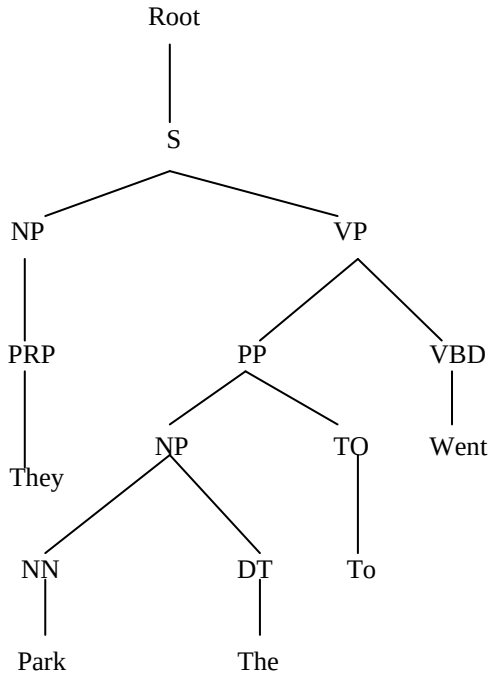

Fig.4 Parse tree for 'They went to the park'

With the differences of word order between English and Indian languages, handling absolutely the reordering problem is necessary. Consider the example sentence “He went to shop”. The following fig. 5 shows the different language word order.

English	He went to shop
Hindi	Vo dhukaanchalgaya
Tamil	Avar kadaikku sendraar
Malayalam	Avankadayil poi

Fig .5 Word order of Indian languages

Where

He : Subject

Went to :Verb

Shop : Object

SVO in source sentence is transformed as SOV in the target sentence as shown in the fig. 6. In this above example the word order of the target sentence is not same as the word order of the source sentence. This reordering is done according to the pre specified rules. These rules are handcrafted by both the linguists and computer programmer. These rules are present inside the database.

He	Subject	avar	Vo	Avan
Went	Verb	sendraar	chalgaya	poi
To Shop	Object	kadai	dhukaan	kadayil

Fig. 6 Target structure of Tamil, Malayalam and Hindi

These rules will be of the pattern source language rule, followed by the target language rule, provided with the transfer information; these are called the transfer rules. This forms the integral part of the translation. The rules are done so far more generic which works for most of the sentences from the English grammatical structures. There are 160 rules in the database covering various cases of the sentences from English language grammar and their corresponding reordered rule structure for Indian languages are formed.

Table 4. Reordering rules

S.No.	CurNode	Source	Target	Transformation
1	ADJP	ADJP PP	PP ADJP	1:2 2:1
2	ADJP	JJ PP	PP JJ	1:2 2:1
3	ADJP	S JJ	S JJ	1:2 2:1
4	ADVP	RB RB	RB RB	1:2 2:1
5	FRAG	NP PP NP-TMP	PP NP-TMP NP	1:2 2:3 3:1 4:4
6	NP	NP, SBAR	SBAR, NP	1:3 2:2 3:1
7	NP	NP NP	NP P	1:2 2:1
8	PP	IN ADVP	ADVP IN	1:2 2:3 3:4 4:1
9	QP	IN CD TO CD	CD TO CD IN	1:1 2:3 3:2 4:4

Some of the rules in the database are as shown in the Table 4. The rules are generated by getting the information from the parser. In the above sentence, the general structure for the source side and target side are shown in fig. 7.

Sentence : He went to shop	
Source Side	$S \rightarrow NPVP$ $VP \rightarrow VBDS$
Target side	$S \rightarrow NPVP$ $VP \rightarrow SVBD$

Fig. 7 Reordering rules for the sentence ‘He went to shop’

E. LEXICAL DICTIONARY

For the purpose of mapping words from one language to another, we require a lexical dictionary. The lexical dictionary contains the word from the source language and its corresponding root word in the target language.

Our dictionary contains three different tables. In case of Indian languages, gender information plays an important role in the formation of a word. The sentence formation depends upon the gender of the subject. Hence it is important to add information about gender in the dictionary. Certain words which are considered to be singular in English are considered as the plural in Indian languages and vice versa. The different categories of those lexicons include nouns, adjective, prepositions, verbs, adverbs and a general category which does not belong to any of these categories.

In this process we simply look up the lexical dictionary in the database and correspondingly take the target words from the dictionary. For example the lookup table for the noun is shown in Table 5.

Table 5. Noun Lexicon Table In Dictionary

SNO	SOURCE	TARGET			CATEGORY	FEATURE
		TAMIL	MALAYALAM	HINDI		
1	shop	kadai	Kada	pesha	NOUN	3SN
2	temple	kovil	ambalam	dihaara	NOUN	3SN
3	temperature	Tatpaveppanilai	thapanila	thapamaana	NOUN	3SN
4	carpenter	thatchar	aashari	padai	NOUN	3SN
5	face	mugam	mugam	mug	NOUN	3SN

F. TRANSLITERATION

We have two conditions that make the lexical look up impossible. They are

- The source word may not have a similar target word in the lexical dictionary table.
- The source language words such as name of person, location, organization or technical terms may not be present in the target language. Hence it becomes mandatory to transliterate these words in order to develop a translated sentence close to the target language providing the complete semantics from the source language.

Transliteration from an informational point of view can be described as *"systematic transliteration is a mapping from one system of writing into another, word by word, or ideally letter by letter."* The transliteration maps the letter of the source word in one to one mapping to the target word letters. The reverse mapping of the transliteration is also possible. Thus the transliteration process involves the following three steps to be done. They are Romanization, Segmentation and Alignment.

The Romanization of the source language word is the process of mapping the letters of the words from the source in the Romanized form of the target language. The mapping can be done by using a map file. The segmentation process converts the English words based on the vowels, consonants, and those segments of words such as *"sh", "bh", "ksh", "ch", "th"* etc… similarly for the Indian language names we segment Romanized words based on its segments of words like the vowels, consonants and segments present for it. The alignment of these words is done properly and is created as the models for learning of the machine learning algorithm. Thus we get the transliterated output in Indian language.

G. MORPHOLOGICAL ANALYZER

Morph analysis is done on the source language side. From the input sentence, we have to extract the following information such as tense information, subject and verb in the sentence. The subject and the verb in the sentence can be easily identified using the dependency relation given by the Stanford Parser.

The Indian language such Tamil are basically agglutinative; PNG markers play an important role in word formation in South Dravidian languages except Malayalam (Discussed in the Section 3). They show unique structural formation in the case of words by the addition of suffixes which bring them different senses and grammatical categories. The addition of these suffixes happens basically after the root words.

Table 6.Derivational morphology for verb (suffixes)

Verb	Suffixes	Distinct form
Vaa(come)	+PRES + " Aan"(1SM)	Varukirran
saapidu(eat)	+FUT+ " gal" (3P)	Saapiduvargal

Verbal categories such as transitive verbs, causative verbs are added along with the following tense information and person, number, gender information's are added. The derivations and inflections are not the identical for every bit of the nouns and verbs. The examples for the derivational morphology of Tamil verb and noun are shown in Table 6 & 7.

Table 7. Derivational morphology for noun (suffixes)

Noun	**suffixes**	**Distinct form**
Naadu(country)	+ " gal" (P)	Naadukal
Manilam(state)	+ "n" + "gal" (P)	Maanilangal

Noun has the following information such as case markers, plural markers and the euphonic increment, whereas the verb needs tense markers and PNG marker to perform the morph generation. The input parameters to the morph generation function are the root word of the main verb, and the tense information extracted. The output is the appropriate form of the verb in the particular tense. Table 4 shows the example of the three cases of verbs.

Table 8. Verb cases for morphological generation

Verb cases	**Output**
Odu+V+FT_3SM	Oduvan
kAdu+N+ACC	kAddai
Maram+N+LOC	maraththil

Thus the sentence of the translation system is being generated at the end of these morph generation processes. This morph generation part adds the verb and noun inflections to the sentence translated.

Consider the example given below; the output of the sentence "He is going to college" after the reordering process will be "**He College to going is**". Here the "**going**" is the verb which is in the continuous tense. Here the root word for that word "going" has to be found which is "**go**". After finding the root word it will be easy to find the meaning in the dictionary. With the help of the tags, the gender ending and tense marker processes are done.

V. PERFORMANCE MEASURES AND EXPREMENTAL RESULTS

We have used the Stanford POS tagger for getting the Parts Of Speech information of the input sentences. So far we have created the word dictionary file which contains almost all commonly used words from English to each Indian language. Gender ending rules for all possible cases were created. The survey of translation was conducted with the proposed hybrid based machine translation by providing a variety of sample English sentences to the online Google translator for the translation from English to Indian languages such as Tamil and Hindi. The validation has been done in accordance with the proposed grammar rules. The proposed system provides the better translation for complex text while compared to Google translators.

For experimental evaluation, the metrics used here is proposed by IBM, called BLEU (Papineni et al. 2002) and the metric urbanized by NIST(NIST, 2002).Both the metrics tried to assess howclose a machine translation is for a set of position translations produced by humans. In the proposed experiments consist of a set of single reference translation presented by the Hansard transcripts.

To calculate the BLEU score, one counts the quantity of n-word fragments (n-grams) in the candidate translations that contains a match in the equivalent reference translations. The n-*gram precision* is this number divided by the total number of n-grams in the candidate translations. BLEU uses a *modified* n-*gram precision*

called p_n. This precision *clips* the count for each n-gram in any candidate translation to avoid it from beyond the count of this n-gram in the best matching reference translation. The N different n-gram precisions are averaged geometrically then multiplied by a *brevity penalty* to discourage short but high-precision translation candidates. This leads to the following formula:

$$\text{BLEU} = e^{\min\left(1-\frac{r}{c},0\right)} . e^{\left(\sum_{n=1}^{N}(1/N)\log p_n\right)}$$

Here, r and c are the total number of words in the reference and candidate translations correspondingly. The brevity penalty $e^{\min\left(1-\frac{r}{c},0\right)}$ is less than one if $c < r$ the candidate translations are shorter than the reference translations on average.

The NIST score is based on related considerations, with three major differences. First, it incorporates an *informed weight* to place more emphasis on infrequent n-grams. Second, it uses arithmetic to a certain extent than geometric average to combine the scores for each n. Third, it uses a brevity penalty that is less sensitive to small variations.

The following formula defines the NIST metric:

$$NIST = e^{\beta log^2\left[\min\left(\frac{c}{r},1\right)\right]} * \sum_{n=1}^{N}\left(\frac{\sum \text{all co} - \text{occurring } w_1 \ldots . . w_n \;\; \text{Info}(w_1 \ldots w_n)}{|allw1 \ldots wnincandidatetranslation|}\right)$$

Here, n-grams up to N = 5 are measured, and it is set such that the brevity penalty is concerning 0.5 for a translation candidate length that is about 2/3 of the reference translation length. The information weights are intended over the reference translation corpus as:

$$Info(w1 \ldots wn) = \log_2\left(\frac{\#\text{occurrences of } w1 \ldots wn - 1}{\#\text{occurrences of } w1 \ldots wn}\right)$$

Both BLEU and NIST scores are sensitive to the number of reference translations. Though both are also sensitive to the number of words in the reference corpus, NIST is much more so because of the language model disguised by the information weights, which are often zero for large n in small corpora.

Comparison of rule based machine, statistical machine translation and hybrid machine translation

Rule-Based MT:

- Syntactically better translations: long distance reordering, agreement.
- Worse lexical selection.
- Performance degradation for unexpected syntactic structures.

Statistical MT:

- Better lexical selection and fluency.
- Structurally worse translations.
- Performance degradation for out-domain texts.

Proposed hybrid machine translation:

- RBMT's grammatical correctness.
- SMT's lexical selection.
- SMT tolerance to unexpected structures.

Our ultimate goal is to provide better accuracy for the translation system. 200 sentences in each language are given to the hybrid English to Indian language machine translation out of which 140 sentences are given incorrect because of syntax and reordering errors. But the same 200 sentences were tested after segmentation and simplification. 115 sentences are translated correctly with simplification. The system provides good results because the input is simplified and segmented before enteringinto the translation process, rule based reordering is performed and finally the errors are corrected statistically. Proposed hybrid machine translation system was evaluated using BLEU and NIST evaluation metric. Table 9 summarizes the comparison of results obtained from the proposed work to rule based and statistical machine translation.

Table 9. Evaluation statistics of proposed hybrid machine translation

Machine translation type	NIST	BLEU
Rule based machine translation	0.8645	0.7258
Statistical machine translation	0.8652	0.7262
Hybrid machine translation	0.8963	0.7923

VI. CONCLUSION

This paper presents an effective methodology for English to Indian languages machine translation. New rules have been added to the proposed system in order to make the translation process more efficient. The results show that significant improvements in the performance there by incorporating certain features such as simplifying & segmenting the input text, adding new reordering rules and adding syntactic and morphological information to the plain corpus. Since all Indian languages relatively rich in terms of morphology and follows the SOV order, the hybrid mechanism that we have implemented should be applicable to English to all Indian languages in general. The system can be further enhanced to other Indian languages by creating new morphological reordering rules according to those languages. The goal of developing such a translation system is to break the language barriers between people and to make the resources available to everyone. And also this proposed system is going to integrate with AutomaticSpeech Recognition (ASR) and Text To Speech (TTS) systems in order to develop a speech to speech translation system for English to Indian languages.

REFERENCES

[1] Dorr, B.J., Jordan, P.W. & Benoit, J.W. 1999. A survey of current paradigms in machine translation. Advances in Computers, 49, 2–68.

[2] Harshawardhan, R., Mridula Sara Augustine and K.P. Soman, 2011. Phrase based English – Tamil Translation System by Concept Labeling using Translation Memory. International Journal of Computer Applications (0975 – 8887), Volume 20– No.3.

[3] Hui Yang, Anne de Roeck, Vincenzo Gervasi, Alistair Willis , and Bashar Nuseibeh, 2011. Analysing anaphoric ambiguity in natural language requirements. Springer-Verlag London Limited.

[4] Koehn, P., Och, F.J. &Marcu, D. 2003. Statistical phrase-based translation. InNAACL '03: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 48–54, Association for Computational Linguistics, Morristown, NJ,USA.

[5] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M.,Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C.,Bojar, O., Constantin, A. &Herbst, E. Moses 2007. Open sourcetoolkit for statistical machine translation. In ACL, The Association for ComputerLinguistics.

[6] Lehmann, T. 1989.. A Grammar of Modern Tamil .Pondicherry Institute of Linguistics and Culture.

[7] Maritz,J.S. 1981.Distribution-Free Statistical Methods, Chapman &Hall. ISBN 0-412-15940-6, page 217.

[8] Myers, Jerome L., Well, Arnold D, 2003. Research Design and Statistical Analysis. Second Edition, Lawrence Erlbaum,pp. 508, ISBN 0-8058-4037-0.

[9] NIST. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.(accessed march 5 2013).

[10] Papineni, K., Roukos, S., Ward, T. & Zhu, W.J. 2001. Bleu: a methodfor automatic evaluation of machine translation. In ACL '02: Proceedings of the40th Annual Meeting on Association for Computational Linguistics, 311–318,Association for Computational Linguistics, Morristown, NJ, USA.

[11] Papineni, K., Roukos, S., Ward, T., and Zhu, W. 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02), pages 311–318.

[12] Poornima C , Dhanalakshmi V, Anand Kumar M and Soman K P, 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. International Journal of Computer Applications (0975 – 8887), Volume 25– No.8.

[13] Rahul.C, Dinunath.K, RemyaRavindran, K.P.Soman, 2009. Rule Based Reordering and Morphological Processing For English-Malayalam Statistical Machine Translation. In the proceedings of the International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp.458-460.

[14] RemyaRajan, Remya Sivan, RemyaRavindran and K.P Soman, 2009. Rule Based Machine Translation from English to Malayalam. In the proceedings of the International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp.439-441.

[15] Samantaray, S., 2007.A Data mining approach for resolving cases of Multiple Parsing in Machine Aided Translation of Indian Languages. In proceedings of Fourth IEEE International Conference on Information Technology, ITNG '07, 2-4, pp.401 – 405, doi:10.1109/ITNG.

[16] Subramanian, K. BBN Technol., Cambridge, Stallard, D. ,Prasad, R. , Saleem, S. , Natarajan, P., 2007. Semantic translation error rate for evaluating translation systems. IEEE Workshop on Automatic Speech Recognition and Understanding.

[17] Wang, Y., Acero, A., and Chelba, C, 2003. Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, US Virgin Islands.

[18] Yamada, K. & Knight, K. 2001. A syntax-based statistical translationmodel. In ACL '01: Proceedings of the 39th Annual Meeting on Associationfor Computational Linguistics, 523–530, Association for Computational Linguistics, Morristown, NJ, USA.

[19] Yang Liu, Dept. of Comput. Sci., Texas Univ., Richardson, TX, Shriberg, E. 2007. Comparing Evaluation Metrics for Sentence Boundary Detection. IEEE International Conference on Acoustics, Speech and Signal Processing,ICASSP.

[20] K. Narayana Murthy, 2000, "Computer Processing of Kannada Language", Seminar on computer and Kannada development, Kannada university, Hampi, 12th October 2000

[21] Antony P J. & Soman K P, 2010. "Kernel Based Part of Speech Tagger for Kannda", International Conference on Machine Learning and Cybernetics, ICMLC 2010, Qingdao, Shandong, China.