A Novel Approach for English to South Dravidian Language Statistical Machine Translation System

Unnikrishnan P
Computational Engineering and
Networking Department,
Amrita Vishwa Vidyapeetham,
Coimbatore, Tamil Nadu, India

Antony P J
Research Scholar
Computational Engineering and
Networking Department,
Amrita Vishwa Vidyapeetham,
Coimbatore, Tamil Nadu, India

Dr. Soman K P
Professor and Head
Computational Engineering and
Networking Department,
Amrita Vishwa Vidyapeetham,
Coimbatore, Tamil Nadu, India

***Abstract*—Development of a well fledged bilingual machine translation (MT) system for any two natural languages with limited electronic resources and tools is a challenging and demanding task. This paper presents the development of a statistical machine translation (SMT) system for English to South Dravidian languages like Malayalam and Kannada by incorporating syntactic and morphological information. SMT is a data oriented statistical framework for translating text from one natural language to another based on the knowledge extracted from bilingual corpus. Even though there are efforts towards building such an English to South Dravidian translation system ,unfortunately we do not have an efficient translation system till now. The first and most important step in SMT is creating a well aligned parallel corpus for training the system. Experimental research shows that the existing methodology for bilingual parallel corpus creation is not efficient for English to South Dravidian language in the SMT system. In order to increase the performance of the translation system, we have introduced a new approach in creating parallel corpus. The main ideas which we have implemented and proven very effective for English to south Dravidian languages SMT system are: (i) reordering the English source sentence according to Dravidian syntax, (ii) using the root suffix separation on both English and Dravidian words and iii) use of morphological information which substantially reduce the corpus size required for training the system. Since the unavailability of full fledged parsing and morphological tools for Malayalam and Kannada languages, sentence synthesis was done both manually and existing morph analyzer created by Amrita university. From the experiment we found that the performance of our systems are significantly well and achieves a very competitive accuracy for small sized bilingual corpora. The proposed ideas can be directly used for other south Dravidian languages like Tamil and Telugu with some minor changes.**

Keywords-SMT; Dravidian languages; parsing; morphology; inflections

I. INTRODUCTION

The term Machine Translation is a standard name for computerized systems responsible for the production of translations from one natural language into another with or without human assistance. It is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. Many attempts are being made all over the world to develop machine translation systems for various languages using rule based as well as statistical based approaches. Literature shows that the rule based machine translation process is extremely time consuming, difficult and failed to analyze accurately a large corpus of unrestricted text. Hence, most modern translation system are based on statistical or at least partly statistical, which allows the system to gather information about the frequency with which various constructions occur in specific contexts. Any statistical approach requires the availability of aligned bilingual corpora which are: large, good-quality and representative.

MT systems can be designed either specifically for two particular languages [1] called bilingual system, or for more than a single pair of languages called multilingual systems. Bilingual system may be either unidirectional, from one source language (SL) into one target language (TL), or bidirectional. Multilingual systems are usually designed to be bidirectional but most bilingual systems are unidirectional. Machine translation (MT) methodologies are commonly categorized as direct [2], transfer, and interlingual. The methodologies differ in the depth of analysis of the source language and the extent to which they attempt to reach a language independent representation of meaning or intent between the source and target languages. These levels of analysis are illustrated with the ‘Vauquois Triangle’ as shown in Fig. 1.

Starting with the shallowest level at the bottom, direct transfer is made at the word level. Moving upward through syntactic and semantic transfer approaches, the translation occurs on representations of the source sentence structure and meaning, respectively. Finally, at the interlingual level, the notion of transfer is replaced with a single underlying representation called the ‘Interlingua’. ‘Interlingua’ represents both the source and target texts simultaneously. Moving up the triangle reduces the amount of work required to traverse the gap between languages, at the cost of increasing the required amount of analysis and synthesis.

Statistical machine translation is a data oriented statistical framework for translating text from one natural language to another based on the knowledge extracted from bilingual corpus. Translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

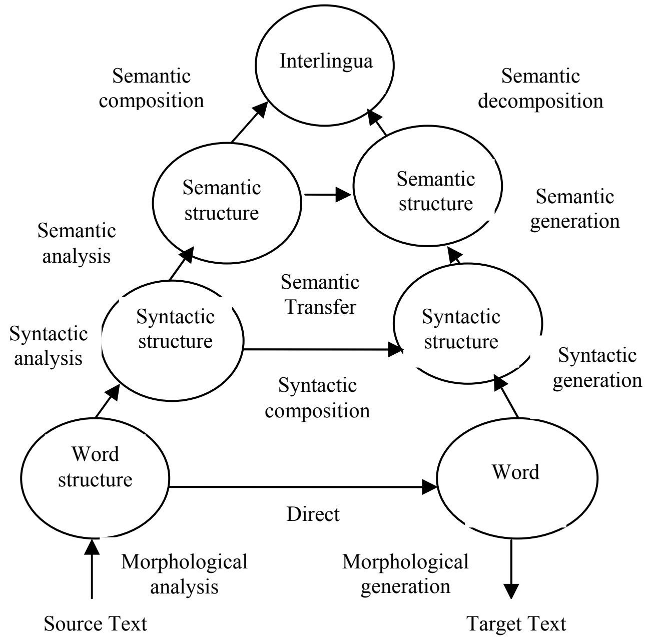


Figure 1. The Vauquois triangle.

The idea behind statistical machine translation comes from information theory. A document is translated according to the probability distribution function indicated by $p(e|f)$. Probability of translate a sentence f in the source language F (for example, English) to a sentence e in the target language E (for example, Malayalam or Kannada).

The problem of modeling the probability distribution $p(e|f)$ has been approached in a number of ways. One intuitive approach is to apply Bayes theorem. That is, if $p(f|e)$ and $p(e)$ indicates translation model and language model respectively, then probability distribution $p(e|f) \infty p(f|e)p(e)$. The translation model $p(f|e)$ is the probability that the source sentence is the translation of the target sentence or the way sentences in E get converted to sentences in F. The language model $p(e)$ is the probability of seeing that target language string or the kind of sentences that are likely in the language E. This decomposition is attractive as it splits the problem into two sub problems. Finding the best translation $\tilde{e}$ is done by picking up the one that gives the highest probability as shown in (1).

$$\tilde{e} = \arg\max_{e \in e^*} p(e \mid f) = \arg\max_{e \in e^*} p(f \mid e) p(e) \qquad (1)$$

Even though phrase based models have emerged as most successful method for SMT they do not handle syntax in a natural way. Reordering of phrases during translation is typically managed by distortion models in statistical machine translation (SMT). But this reordering process is entirely not satisfactory especially for language pairs that differ a lot in terms of word-order. In the proposed project the problem of structural differences between source and target languages are successfully overcome with a reordering task. We have also proven that with the use of morphological information, especially for morphologically rich languages like Malayalam and Kannada, the training data size can be much reduced with an improvement in performance.

II. Literature Survey

A first public Russian to English [3] machine translation system was presented at the University of Georgetown in 1954 with a vocabulary size of around 250 words. Since then, many research projects were devoted to machine translation during the late 1950s. However, as the complexity of the linguistic phenomena involved in the translation process together with the computational limitations of the time were made apparent, enthusiasm faded out quickly. Also the results of two negative reports namely 'Bar-Hillel' and 'AL- PAC' had a dramatic impact on machine translation research by that decade.

During the 1970s, the focus of machine translation activity switched from the United States to Canada and to Europe, especially due to the growing demands for translations within their multicultural societies. 'Mateo', a fully-automatic system translating weather forecasts had a great success in Canada. Meanwhile, the European Commission installed a French–English machine translation system called 'Systran'. Other research projects, such as 'Eurotra', 'Ariane' and 'Susy', broadened the scope of machine translation objectives and techniques. The rule-based approaches emerged as the right way towards successful machine translation quality. Throughout the 1980s many different types of machine translation systems appeared and the most prevalent being those using an intermediate semantic language such as the 'Interlingua' approach.

Lately, various researchers have shown better translation quality with the use of phrase translation. Most competitive statistical machine translation systems such as the CMU, IBM, ISI, and Google etc. used phrase-based systems and came out with good results.

In the early 1990s, the progress made by the application of statistical methods to speech recognition introduced by IBM researchers was purely-statistical machine translation models [3]. The drastic increment in computational power and the increasing availability of written translated texts allowed the development of statistical and other corpus-based machine translation approaches. Many academic tools turned into useful commercial translation products, and several translation engines were quickly offered in the World Wide Web.

Today, there is a growing demand for high-quality automatic translation. Almost all the research community has moved towards corpus-based techniques, which have systematically outperformed traditional knowledge-based techniques in most performance comparisons. Every year more research groups embark on SMT experimentation and a regained optimism as regards to future progress seems to be shared among the community.

A. Related works

In 1999 Franz Josef Och, Christoph Tillmann, and Hermann Ney propose improved alignment models for statistical machine translation [4]. In 2000 Durgesh Rao, Kavitha Mohanraj, and Jayprasad Hegde present a practical framework for the syntactic transfer of compound-complex sentences from English to Hindi in the context of a transfer-based Machine Assisted Translation (MAT) system [5]. Kenji Yamada and Kevin Knight in 2001 present a syntax-based statistical translation model. Their model transforms a source-language parse tree into a target-language string by applying stochastic operations at each node [6]. In 2002 Daniel Marcu and William Wong present a joint probability model for statistical machine translation, which automatically learns word and phrase equivalents from bilingual corpora [7]. Philipp Koehn, Franz Josef Och and Daniel Marcu propose a new phrase-based translation model and decoding algorithm in 2003 that enables to evaluate and compare several, previously proposed phrase-based translation models [8]. Franz Josef Och in 2003 analyzes various training criteria which directly optimize translation quality [9]. These training criteria make use of recently proposed automatic evaluation metrics. B. Pang, K. Knight, and D. Marcu in 2003 describe a syntax-based algorithm that automatically builds Finite State Automata from semantically equivalent translation sets [10]. Kenji Imamura, Hideo Okuma, Eiichiro Sumita in 2004 presents a practical approach to statistical machine translation based on syntactic transfer [11]. In 2004 I. Dan Melamed explains generalizations of ordinary parsing algorithms that allow the input to consist of string tuples and/or the grammar to range over string tuples [12]. In 2005 Collins and Koehn describe a method for incorporating syntactic information in statistical machine translation systems [13]. Sonja Nieben and Hermann Ney , in 2004 introduce sentence-level restructuring transformations which aim at the assimilation of word order in related sentences [14]. Maja Popovic and Hermann Ney in their work in 2006 concluded that the performance of a statistical machine translation system depends on the size of the available task-specific bilingual training corpus [15]. Michael Collins, Philipp Koehn, and Ivona Kucerova in 2006 describe a method for incorporating syntactic information in statistical machine translation systems [16]. Maja Popovic and Hermann Ney in 2006 investigated new possibilities for improving the quality of statistical machine translation (SMT) by applying word reordering of the source language sentences based on Part-of-Speech tags [17]. Marcu, D., Wang, W. and Echihabi in 2006 introduced SPMT, a new class of statistical Translation Models that use Syntactified target language Phrases [18]. The SPMT models outperform a state of the art phrase-based baseline model by 2.64 Bleu points on the NIST 2003 Chinese-English test corpus and 0.28 points on a human based quality metric that ranks translations on a scale from 1 to 5. Nizar Habash in 2007 describes an approach to automatic source-language syntactic pre-processing in the context of Arabic-English phrase-based machine translation [19]. Ibrahim Badr, Rabih Zbib, and James in 2008 show that morphological decomposition of the Arabic source is beneficial, especially for smaller-size corpora, and investigate different recombination techniques [20]. They also report on the use of Factored Translation Models for English-to-Arabic translation. 1n 2008, Ananthakrishnan Ramanathan and Pushpak Bhattacharyya reports their work on incorporating syntactic and morphological information for English to Hindi statistical machine translation [21].

III. SOUTH DRAVIDIAN LANGUAGES

Among the four major South Dravidian languages such as Kannada, Malayalam, Tamil and Telugu are having almost 40, 35, 70 and 71 million speakers respectively [22]. These languages have their own independent scripts and long documented histories. Verbs have a negative as well as an affirmative voice. Gender classification is made on the basis of rank instead of sex, with one class including beings of a higher status and the other beings of an inferior status. Nouns are declined, showing case and number. In South Dravidian languages a great use is made of suffixes with nouns and verbs. Also all these four Dravidian languages have their own alphabets, related to the Devanagari alphabet used for Sanskrit. Even though Malayalam and Kannada are languages of rich in historical literary, they are resource poor when viewed through the prism of computational linguistics [23]. In this paper most of the descriptions are based on Malayalam and Kannada languages.

According to the most dependable evidences available to us, the Malayalam has its root to the 10th century and literature is at least a thousand years old. Malayalam has drawn influence from both Indian and foreign languages, such as Tamil, Sanskrit, Prakrit, Pali, Hebrew, Hindi, Urdu, Arabic, Persian, Syriac, Portuguese, Dutch, French and English. [24]. Out of 1652 mother tongues being spoken in India, Kannada has been estimated to be over 2, 500 years old and the third oldest Indian language after Sanskrit and Tamil. The Kannada language is one of the four major Dravidian languages of South India. It is the state language of Karnataka and is spoken by about 20 million people. It has a long linguistic of about 1,500 years and had a continuous literature for over 1,200 years. Literature shows that computationally a very little research has been done in Kannada natural language processing.

A. Structure of Malayalam and Kannada Languages

Malayalam and Kannada are highly agglutinative language with three gender forms namely masculine, feminine and neutral or common. Singular and plural are the two number forms that interestingly shows inflection based on the gender, number and tense of the commodity of reference among other factors.

Both Malayalam and Kannada languages are ‘Left branching language’, in which verbs are usually at the end of the sentence and have post positions instead of prepositions. Hence adjectives, genitive and relative clauses precede their head nouns in a sentence. Past and non-past are the two broad types of tenses in Malayalam and Kannada languages. ‘Mood’ is another important feature and is associated with statements of fact versus possibility, supposition, etc. There are four different moods that are expressed are: infinitive, imperative, affirmative and negative. Also these two languages have some additional ‘modal’ forms such as: indicative, conditional, optative, potential, monitory and conjunctive.

The noun phrase (NP) of these two languages is simple and has adjectives derived from nouns or verbs and nouns of various sorts that take case endings and post positions. In some cases NP may contain pronouns, numerals, color terms, deictic particles such as ‘ഇത്’, 'ಇದು' (this), ’അത്’, 'ಅದು' (that), ’ഏത്’, 'ಯಾವುದು' (which), etc., and quantifiers like ’കുറെ’, ‘ತುಂಬಾ’ (many), ’കുറച്ച്’, 'ಸ್ವಲ್ಪ', (some) etc. NP may consist of nominal head or pronoun and may be followed by modifiers. Syntactically noun phrases are identified by their potential to act as subjects, direct objects, indirect objects and compliment of postpositional phrases.

Word order plays an important role in positional languages like English and normally follow right-branching with Subject-Verb-Object orders. Unlike English language Malayalam and Kannada are syntax of relatively free word order language [25]. 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. This can be easily illustrated with the example 'India defeated Pakistan in Lahore' as shown in table 1.

TABLE I. WORD ORDER IN MALAYALAM AND KANNADA LANGUAGES

Case	Malayalam	Kannada
Case 1	ഇന്ത്യ പാകിസ്താനെ ലാഹോറിൽ തോൽപിച്ചു.	ಭಾರತವು ಪಾಕಿಸ್ತಾನವನ್ನು ಲಾಹೋರಲ್ಲಿ ಸೋಲಿಸಿತ್ತು.
Case 2	പാകിസ്താനെ ഇന്ത്യ ലാഹോറിൽ തോൽപിച്ചു.	ಪಾಕಿಸ್ತಾನವನ್ನು ಭಾರತವು ಲಾಹೋರಲ್ಲಿ ಸೋಲಿಸಿತ್ತು.
Case 3	ഇന്ത്യ ലാഹോറിൽ പാകിസ്താനെ തോൽപിച്ചു	ಭಾರತವು ಲಾಹೋರಲ್ಲಿ ಪಾಕಿಸ್ತಾನವನ್ನು ಸೋಲಿಸಿತ್ತು.
Case 4	ലാഹോറിൽ ഇന്ത്യ പാകിസ്താനെ തോൽപിച്ചു	ಲಾಹೋರಲ್ಲಿ ಭಾರತವು ಪಾಕಿಸ್ತಾನವನ್ನು ಸೋಲಿಸಿತ್ತು.

Even though all the sentences above certainly do not mean exactly the same thing, they are all equivalent as far as the functional structure of the sentence is concerned. In all the cases, the subjects are 'ഇന്ത്യ' (inthya) and 'ಭಾರತ' (bhArata), the objects are 'പാകിസ്താൻ' (pAkisthAn) and 'ಪಾಕಿಸ್ತಾನ' (pAkistAna) and the locative is 'ലാഹോർ' (lAhOr), 'ಲಾಹೋರ್'. From the above example, it is clear that word order does not determine the functional structure in South Dravidian languages and permits scrambling. But normally South Dravidian languages follow Subject-Object-Verb orders in contradiction with English language.

B. *Complexity and Ambiguity*

The highly agglutinative languages like Malayalam and Kannada, nouns and verbs get inflected. 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 [26]. This leads to the complexity in bilingual machine translation. A noun may be categorized as common, proper or compound. Similarly, verb may be finite, infinite, gerund or contingent. 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 is the another important issue that have to be carefully analyse while designing a machine translation system. For example, Malayalam word 'കാലി' (kAli) and the Kannada word 'ಬತ್ತಿ' (batti) in the following sentences in table 2 gives different parts of speech.

TABLE II. AMBIGUITY IN MALAYALAM AND KANNADA LANGUAGES

Case	Malayalam	Kannada
Case 1	അവന് കാലി തൊഴുത്തില് ജനിച്ചു (avan kAli tozhuthil janichu)	ಸೀತೆ ದೀಪದ ಬತ್ತಿ ಬದಲಿಸಿದಳು. (Seete deepada batti badalisidaLu)
Case 2	പോക്കറ്റ് കാലി ആയി (pOcket kAli Ayi)	ಬಾವಿಯ ನೀರು ಬತ್ತಿ ಹೋಯಿತ್ತು. (baaviya neeru batti hooyittu)

The words 'കാലി' (kAli) and 'ಬತ್ತಿ' (batti) are nouns in the first case whereas in the second case these words act as verbs.

C. *Person-Noun-Gender (PNG) and Tense Markers*

The PNG and the tense marker concatenated to the verb stems are the two important aspect of verb morphology in South Dravidian languages. 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 determines the agreement marker of the verb. Usually the South Dravidian languages verbs follow the regular pattern of suffixation. The table 3 shows the various PNG suffixes that can be attached to be any Kannada verb root word.

TABLE III. PNG- SUFFIXES IN KANNADA

			PNG Suffix			
P	N	G	Present	Future	Past	Contingent
1st	S	M/F	ಏನೆ (Ene)	ಎನು, ಎ (enu, e)	ಎನು, ಎ (enu, e)	ಏನು (Enu)
	P	M/F	ಏವೆ (Eve)	ಏವು (Evu)	ಎವು (Evu)	ಏವು (Evu)
2nd	S	M/F	ಈ, ಈಯೆ (I, Iye)	ಈ, ಇಯೆ (I, iye)	ಇಯ (izha)	ಈಯ (Izha)
	P	M/F	ಈರಿ (Iri)	ಈರಿ (Iri)	ಇರಿ (iri)	ಈರಿ (Iri)
3rd	S	M	ಆನೆ (Ane)	ಆನು (Anu)	ಅನು (anu)	ಆನು (Anu)
	S	F	ಆಳೆ (Ale)	ಅಳು (aLu)	ಅಳು (aLu)	ಆಳು (ALu)
	P	M/F	ಆರೆ (Are)	ಆರು (Aru)	ಅರು (aru)	ಆರು (Aru)
	S	N	ಇದೆ (ide)	ಉದು(udu)	ಇತು (itu)	ಈತು (Ittu)
	P	N	ಇವೆ (ive)	ಅವು (Avu)	ಅವು (avu)	ಆವು (Avu)

P: Person; N: Number; G: Gender

S: Singular; P: Plural; M: Masculine; F: Feminine; N: Neuter

Depends on the noun case associated with NNP, the PNG marker in the VF may change. That is, if NNP indicate feminine then the PNG marker is ಅಳು 'aLu'. If NNP indicate a name of a respected person then the PNG marker is ಅರು 'aru' instead of ಆನು 'anu' or ಅಳು 'aLu'. But in case Malayalam, regardless of noun case, there is no PNG marker followed to the tense marker associated with the verb.

In both cases regardless of the type of verb paradigms, all the verb words use the same present and future tense markers.

But all the South Dravidian languages uses different past tense markers based on the types of verb paradigms. The table 4 shows the different tense markers that are used in Kannada and Malayalam languages.

TABLE IV. TENSE MARKERS IN MALAYALAM AND KANNADA

	Tense Markers	
Tense	Malayalam	Kannada
Present	‘ഉന്നു’ (unnu)	ಉತ್ತ್ (Utt)
Future	‘ഉം’ (um)	ಉವ್ (Uv)
Past	ഇ (i,), ന്നു (nnu), ഞ്ഞു(nju), റ്റു(tu), ട്ടു(ttu), ക്ക (ccu), കു(cu), ന്റു(ntu), ണ്ടു(NTu)	ತ್ತ್(tt), ಂತ್(Mt).,ತ್(t), ದ್(d), ದ್ದ್ (dd), ಇದ್ (id), ಂದ್(Md), ಡ್(D),ಟ್(T), ಕ್ಕ್(kk), ಂಡ್(MD)

IV. EXPERIMENTAL FRAMEWORK

The proposed English to Dravidian languages like Malayalam and Kannada SMT system successfully found the solution for the following major challenges: i) The difference in the word order of English and Malayalam (or Kannada) ii) Morphological differences between English and Malayalam (or Kannada) and iii) Availability of parallel corpora.

A well organized proper bilingual corpus is the most important thing for an efficient SMT system. From our experiment we inferred that if we are using the English-Malayalam (or Kannada) corpus as such to the SMT system for training, the translation system results very poor performance. One obvious reason for the poor translation is the linguistic distance between the source and target languages. The most significant difference is in the word order or chunk order of source and target languages. The underlying structural differences between the source and target languages which forms a major weighting factor for the low translation quality and manifest themselves as a relatively poor translation. In this work the problem of structural differences between source and target languages are successfully overcome with reordering task. In our bilingual corpus we reordered the sentences to make the order of corresponding chunks to be same. We reordered the English sentences according to Malayalam (or Kannada) sentence structure by including some structural or syntactic information to our SMT system. Reordering was performed by retrieving the structural information of English by using an English parser and changes the structure according to the Malayalam (or Kannada) word order.

The second challenge that really matter in the SMT system is the morphological difference between English and Malayalam (or Kannada). If an SMT system considers different morphological forms of a word as independent entities, a crucial source of information is neglected. South Dravidian languages like Malayalam and Kannada are morphologically very rich than English. From our experiment we have proven that with the use of morphological information, especially for morphologically rich languages like Malayalam and Kannada, the requirement for training data can be much reduced.

The third major challenge specific to all Indian languages especially for South Dravidian languages like Malayalam and Kannada are the availability of parallel corpora. The translation quality improves with amount of parallel corpora which should be: i) large ii) good quality and iii) representative.

A. *Why Translation Need Data Pre-processing*

The structural difference between English and Malayalam (or Kannada) greatly affects the performance of the translation system. This can be easily illustrated with the following example. Consider a simple English sentence ‘I am going to school’. The corresponding Malayalam and Kannada sentences are ‘ഞാൻ സ്കൂളിലേക്ക പോയികൊണ്ടിരിക്കുന്നു ‘ (njAn schooLilekku poyikkondirikkunnu) and ‘ನಾನು ಶಾಲೆಗೆ ಹೋಗುತ್ತಿದ್ದೇನೆ’ (nAnu shAlege hOguttiddEne). The Fig. 2 shows the structural difference between source and target which demands pre-processing for improving the performance in a greater extent.

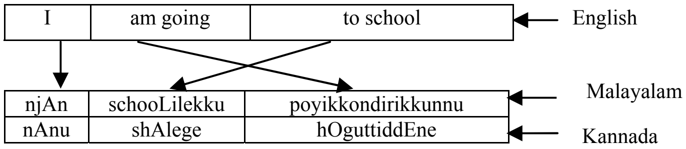


Figure 2. Structural difference between English and Dravidian languages.

The structural difference between the sentences is directly proportional to the size or length of the sentences. Reordering of phrases during translation is typically managed by distortion models. But experiments show that they have not entirely satisfactory especially for language pairs that differ much in terms of word-order. To get over this drawback we have used a pre-processing approach, by reordering the English sentences in the training and test corpora before the SMT system kicks in. This reduces, and often eliminates, the ‘distortion load’ on the phrase-based system.

B. *Tools Used*

In the proposed English to Dravidian SMT system we have used various natural language processing tools. Each and every tool has its own functionalities and used for various purposes. The description of these tools is as follow.

1) Language model(LM), Translation model(TM), Decoder and BLEU: The statistical machine translation system requires three prime components namely Language model, Translation model and Decoder. The open source tools such as, Stanford research institute language model (SRILM) and GIZA++ were used for creating language and transliteration model. Another tool called MOSES, a beam search decoder was used for English to Malayalam (or Kannada) translation. Finally the model was evaluated using BLEU, an evaluator for machine translation commonly used evaluator for SMT.

a) Creating language model using SRILM: SRILM is a toolkit for building and applying various statistical language models. The main objective of SRILM is to support language model estimation and evaluation. Estimation creates a model from training data. Evaluation compute the probability of a test corpus for which conventionally expressed as the test set perplexity. Normally SRILM performs the following three important functions: i) generation of n-gram count file from the monolingual target language corpus, ii) training the language from the n-gram count file and iii) calculation of the test data perplexity using the trained language model. It requires huge well organized monolingual target language corpus such as Malayalam or Kannada as training data. The functional block diagram of SRILM is shown in Fig. 3.

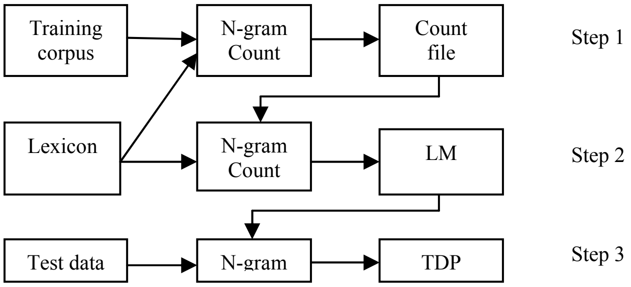


Figure 3. Steps for developing language model.

SRILM are based on N-gram statistics such that the probability distribution *P(e)* for a sentence containing the word sequence w1w2 . . .wn can be written as shown in (2).

$$P(e)=P(w1)P(w2|w1)P(w3|w1w2)..P(wn|w1w2\ ..wn-1) \qquad (2)$$

b) Training of statistical translation models using GIZA++: GIZA++ is an extension of the program GIZA which was developed by the Statistical Machine Translation team. GIZA++ implements IBM-4 (also IBM-5) alignment model with a dependency of word classes and also implement an alignment model based on Hidden Markow model (HMM) using Baum-Welch training and Forward-Backward algorithm. Giving a bilingual parallel corpus to GIZA++ which in turn using an unsupervised learning algorithm called Expectation–Maximization algorithm, the translation probabilities are computed.

The role of the translation model *P(f|e)* is to find the probability of the source sentence *f* given the translated sentence *e*. An aligned parallel corpus of the source and target language sentences are used for training the translation model. The n-gram translation model uses data driven approach to generate probabilistic transformation rules. The value of *P(f|e)* is estimated based on the n-gram statistics obtained by using one to one mapping between the translation units of the source language and target language sentences. Alignment between the sentences at the word level tells the manner in which a word in a sentence *e* is translated to a word in *f*. The probability of *f* given *e* denoted as *P(f|e)* is calculated as in (3).

$$P(f|e)=\sum_M P(f,\ M|e)=\sum_M P(f|M,\ e)\ P(M|e) \qquad (\ 3)$$

where *M* is the match type defined as a pair of translation unit lengths for the source and target languages. When the summation criterion in *P(f|e)* is approximated into maximization, the computational complexity is reduced as in (4).

$$P(f|e) = \max \sum_M P(f|M,\ e)\ P(M) \qquad (4)$$

c) MOSES: The translation of English to Malayalam (or Kannada) was performed with the decoder MOSES. MOSES trains automatically the translation models for any language pair based on the parallel corpus. MOSES is a Beam-Search Decoder for Factored Phrase-Based SMT models. Beam search is an efficient algorithm that finds the highest probability translation among the exponential number of choices. The phrase-based approach allows the translation of short text chunks and the words that may have factored representation and this process is the state-of-the-art in SMT. MOSES also performs the decoding of confusion networks. It features novel factored translation models, which enable the integration of linguistic and other information at many stages of the translation process [27]. Using the learned parameters the decoder performs the translation using the trained model that was created already. The decoder uses the modified Viterbi and A* algorithms to search for highest probability translation that satisfies the condition as in (1).

Moses features novel factored translation models, which enable the integration linguistic and other information at many stages of the translation process. In Moses a phrase-based translation model consists of i) a phrase translation table called *phrase-table* and ii) a configuration file for the decoder called *moses.ini*. The key to good translation performance is having a good phrase translation table. In addition some tuning can be done with the decoder. The most important is the tuning of the model parameters.

The probability cost that is assigned to a translation is a product of probability costs of four models such as phrase translation table, language model, reordering model, and word penalty. Each of these models contributes information over one aspect of the characteristics of a good translation:

- The *phrase translation* table ensures that the English phrases and the Malayalam (or Kannada) phrases are good translations of each other.
- The *language model* ensures that the output is fluent target language like Malayalam (or Kannada).
- The *distortion model* allows for reordering of the input sentence, but at a cost: The more reordering, the more expensive is the translation.
- The *word penalty* provides means to ensure that the translations do not get too long or too short.

The basic reordering model implemented in the decoder is fairly weak. Reordering cost is measured by the number of words skipped when English phrases are picked out of order. Total reordering cost is computed as in (5).

$$D(e,f) = -\ \Sigma_i\ (d\_i) \qquad (5)$$

where *d* for each phrase *i* is defined as d = abs (last word position of previously translated phrase + 1 - first word position of newly translated phrase).

d) Bilingual Evaluation Understudy: Bilingual Evaluation Understudy (BLEU) is an automatic evaluation method, in which metric is based on n-gram co-occurrence based measure. The intrinsic quality of machine translation output is judged by comparing its n-grams with reference translations by humans. Dravidian languages are morphologically rich with lot of suffixes results only rough translations. Generally BLEU is not much appropriate for rough translations and there for not an efficient evaluating tool for English to Dravidian language translation.

2) The Stanford statistical parser: In the proposed project the well known Stanford parser was used to parse the English sentence. The output syntactic information produced by the parser was used to reordering the English sentence. The parser can read various forms of plain text input and can output various analyses formats, including part-of-speech tagged text, phrase structure trees, and grammatical relations (typed dependency) format.

3) Roman to Unicode and Unicode to Roman converter: A well organized aligned bilingual corpus of English and Malayalam (or Kannada) was created by including all types of sentences. SMT support only Roman character code but Dravidian language like Malayalam and Kannada does not support this code format and support only Unicode character. Unicode or officially called the Unicode Worldwide Character Standard is an entirely new idea in setting up binary codes for text or script characters. It is a system for "the interchange, processing, and display of the written texts of the diverse languages of the modern world. A Unicode character set that encompasses all of the world's living languages and is the basis of most modern software internationalization. Unicode is an industry standard whose goal is to provide the means by which text of all forms and languages can be encoded for use by computers. So in order to map from Unicode to Roman and vice versa we have created and used two different mapping files. The table 5 below shows the example for Romanization.

TABLE V. ROMANIZATION

English word	Malayalam		Kannada	
	Malayalam word	Romanized Malayalam	Kannada word	Romanized Kannada
karnataka	കർണാടക	kaRNATaka	ಕರ್ನಾಟಕ	karnATaka
kerala	കേരളം	kEraLam	ಕೇರಳ	kEraLa

4) Morphological analyzer and generators: The role of morphology is very significant in the field of NLP, as seen in applications like machine translation (MT), question-answering (QA) system, IE, IR, spell checker, lexicography etc. So from a serious computational perspective the availability of a morphological generator and analyzer for a language is important. Morphology is the study of word formation and structure. It studies how words are put together from their smaller parts called morphemes and the rules governing this process. Morphological analysis is the process of splitting the word to morphemes. Here our aim is to get the root words for given sentences. The roll of morphological generator is just reverse of that analyzer. It generates a meaning full word from one or more morphemes.

a) English morphological analyzer: We have used a Stanford parser as an English morphological analyzer to analyze the English sentences. To get the root word for a given word it is necessary to give both the word and the corresponding part of speech (POS) tag as an input to the Stanford parser. This is illustrated with examples in table 6.

TABLE VI. EXAMPLE FOR EXTRACTING ENGLISH ROOT WORDS

Input (word/POS tag)	Output (root word)
went/VBD	go
Going/VBG	go
Tables/NNS	table

b) Malayalam and Kannada morphological analyzers: In case of Malayalam and Kannada, we are not only interested on the root word but also its inflection. But the fact that full fledged morphological analyzer and generators are currently not available for Malayalam and Kannada. In the proposed project, an SVM based statistical morphological analyzer and generator developed by AMRITA university was used to extract the root word and inflections attached with a particular word in some extent. The table 7 illustrates some simple examples for noun (first two rows) and verb (third row) words analyzing for Malayalam and Kannada.

TABLE VII. MORPHOLOGICAL ANALYSIS EXAMPLES

Malayalam		Kannada	
Input	Output	Input	Output
ആനകള് (AnakaL~)	ആന+കള് (Ana+kaL~)	AnegaLu (ಆನೆಗಳು)	ಆನೆ+ಗಳು (Ane+gaLu)
മേശയുടെ (mESayute)	മേശ+ഉടെ (meSa+ude)	ಮೇಜಿನ (mEjina)	ಮೇಜು+ಇನ (mEju+ina)
പോകും (pOkum)	പോകു+ഉം (pOku+um)	ಹೋಗುತ್ತೇನೆ hOguttEne)	ಹೋಗು+ಉತ್ತ್+ಏನೆ (hOgu+utt+Ene)

c) Malayalam and Kannada morphological generators: Morphological generation is the reverse process of morphological analysis. The function of morphological generator is to combine the constituent morphemes to get the actual word. In this project we have used an SVM based statistical morphological generators to perform this task. The table 8 illustrates some simple examples for noun (first two rows) and verb (third row) words generation for Malayalam and Kannada.

TABLE VIII. MORPHOLOGICAL GENERATION EXAMPLES

Malayalam		Kannada	
Input	Output	Input	Output
ആന+കള് (Ana+kaL~)	ആനകള് (AnakaL~)	ಆನೆ+ಗಳು (Ane+gaLu)	AnegaLu (ಆನೆಗಳು)
മേശ+ഉടെ (meSa+ude)	മേശയുടെ (mESayute)	ಮೇಜು+ಇನ (mEju+ina)	ಮೇಜಿನ (mEjina)
പോകു+ഉം (pOku+um)	പോകും (pOkum)	ಹೋಗು+ಉತ್ತ್+ಏನೆ (hOgu+utt+Ene)	ಹೋಗುತ್ತೇನೆ (hOguttEne)

5) Transfer rule file: Transfer rule file is used for maintain various information necessary for transforming source structure to the target structure. The first field in the transfer rule file contains all the possible productions used to generate the English sentences. The second field contains new productions corresponding to production in first field which gives information about how source sentence has to be reorderd according to target sentence structure. The third field is called the transfer links which describes about the changes in the target structure with respect to the source structure. Table 9 indicates some entries in Transfer rule file for the sentence ‘He is speaking Malayalam’. After applying these transfer rules, the sentence will convert into “He Malayalam speaking is”.

TABLE IX. EXAMPLE OF TRANSFER RULES

Productions rules for normal English sentence	Productions rules for reorderd English sentence	Transfer links
S ⟶ NP VP	S ⟶ NP VP	0:1,1: 0
VP ⟶ VBZ VP	VP ⟶ VP VBZ	0:1, 1:0
VP ⟶ VBG NP	VP ⟶ NP VBG	0:1, 1:0

C. Implementation

The architecture of the proposed translation system is shown in figure in Fig. 4.

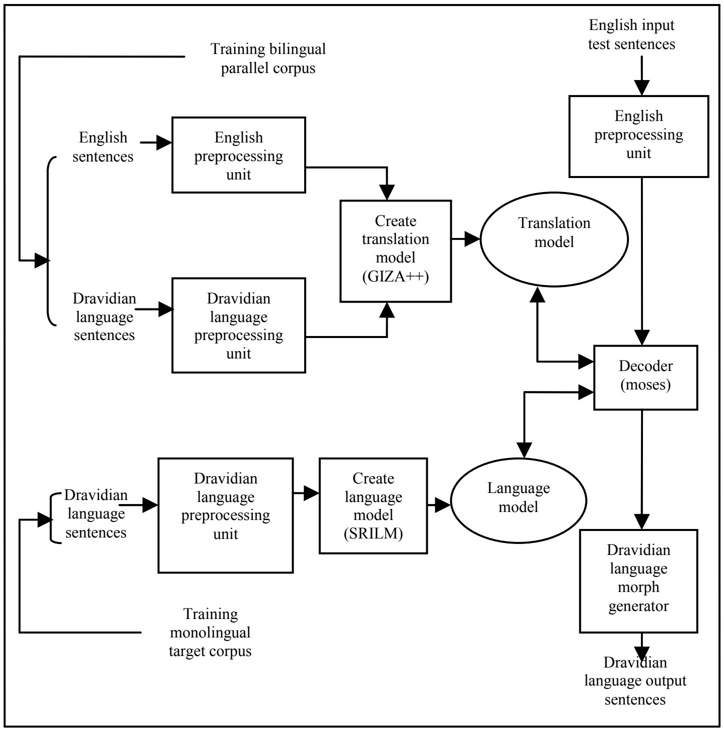

Training bilingual parallel corpus
English input test sentences
English sentences
English preprocessing unit
English preprocessing unit
Create translation model (GIZA++)
Translation model
Dravidian language sentences
Dravidian language preprocessing unit
Decoder (moses)
Dravidian language sentences
Dravidian language preprocessing unit
Create language model (SRILM)
Language model
Dravidian language morph generator
Training monolingual target corpus
Dravidian language output sentences

Figure 4. Proposed English to Dravidian Language SMT System.

The various tools that are used in the proposed system are explained in the previous section. The following sub sections describe the working procedure of the proposed translation system.

1) Pre-processing of source sentences: The main idea behind the improvement in performance of the proposed translation system is pre-processing of data. The Fig. 5 shows the functional block diagram of pre-processing system. Pre-processing of English sentence is a two steps process as: i) Reordering the English sentence and ii) Adding morphological information.

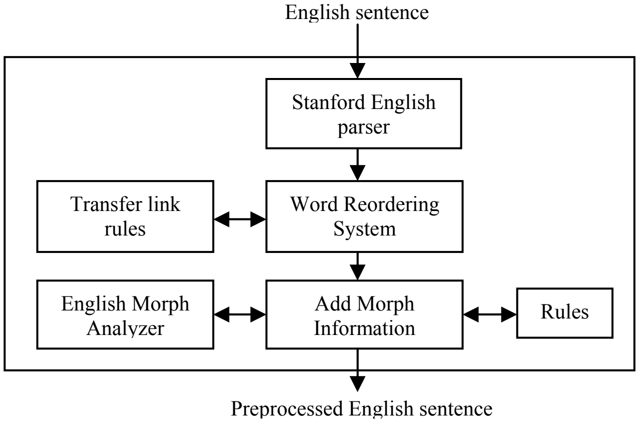


Figure 5. Pre-processing of English Sentences.

a) Reordering of English sentence: The main intension of reordering of English sentence was to match the corresponding source and target phrases same in the sentences. Reordering of English sentence consists of the following steps:

- Get the structural information of English sentence.
- Examine the syntactic structure of sentence.
- Change the structure of English sentence to match the Malayalam (or Kannada) sentence structure.

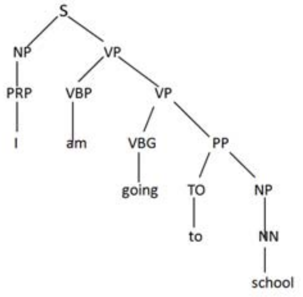


Figure 6. Example: Tree structure before reordering.

A Stanford English parser was used to exploring the structure of the English sentence. The corresponding POS tags for the constituents in a sentence were extracted from the output tree structure of Stanford parser. In the next step, identified the production rules used by the parser to make the parse tree structure. Finally the Transfer link rule file was used to change the structure of English sentence according to Malayalam (or Kannada) chunk order and these changes were applied on the tree. The reordering steps are illustrated with the following example sentence ‘I am going to school’. In the first step, the sentence was given to the Stanford parser which in turn produces the parse tree structure of the corresponding sentence as shown in Fig. 6.

In the next step identified the following productions or rules used by the parser to make the parse tree structure.

S->NP VP
VP->VBG VP
VP->VBG PP
PP->TO NP

Finally the transfer rule file was used to change the structure of English sentence according to Malayalam (or Kannada) chunk order as shown below.

S->NP VP
VP->VP VBG
VP->PP VBG
PP->NP TO

After applying the transfer rules, the corresponding parse tree also changes as shown in Fig. 7.

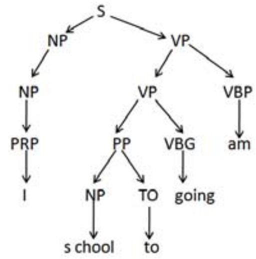
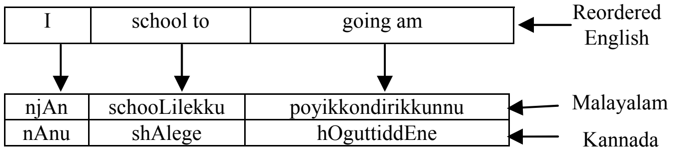


Figure 7. Example: Tree structure after reordering.

After reordering the English sentence ‘I am going to school’ became ‘I school to going am*’* which reduces the structural difference between source and target sentences as shown in Fig. 8. As a result reordering reduces the ‘distortion load’ on the phrase-based system. In our English to Malayalam and English to Kannada SMT systems, Transfer rule files consists of more than 100 entries each, which are capable of handling almost all simple sentences of length even more than ten words.



I	school to	going am

njAn	schooLilekku	poyikkondirikkunnu
nAnu	shAlege	hOguttiddEne

Figure 8. Structural similarities after reordering.

b) Incorporating Morphological Information: Generally SMT system considers different morphological forms of a word as independent entities which in turn increase the size of the corpus. From the experiment it was observed that with the use of morphological information, the requirement of training data size can be substantially reduced. This morphological information plays an important role in the SMT based translation system especially for English to morphologically rich Dravidian languages like Malayalam and Kannada. For example the words “table” and “tables” are considered as entirely different entities in the general SMT systems. But by adding morphological information, the SMT system can identify these words are different form of same word “table”.

A morphological analyzer available with Stanford parser was used to get all the root words of each and every sentence. Input to this morphological analyzer was a combination of word and its POS tag sequences of the reordered sentence as below.

Input**: ‘**I/FW school/NN to/TO going/VBG am/VBP’

The morphological analyzer produces the following result for the given input.

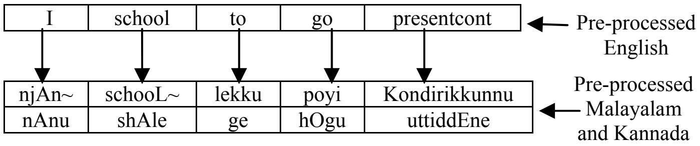
Output**:** “I/FW school/NN to/TO go/VBG be/VBP”

As a further improvement in translation performance, we have developed and applied various rules for different types of sentences. As a result the sentences are modified in a more generalized form suitable to reduce the corpus size. For example the result of morphological analyzer for the example sentence is again modified into the following form.

‘I school to go presntcont’

In the above case the word “presentcont” is corresponding to the Malayalam and Kannada present continuous marker such as “kondirikkinnu” and “uttiddEne” respectively. Similarly rules have been written for various types of simple sentences for all twelve tenses, their negative and question forms.

2) Pre-processing of Target Sentences: On the other side morphological analysis of sentence was the only pre-processing task required for target language. All the words for a given sentence were splits into morphemes of root words and its inflections as explained in the previous section. Fig. 9 illustrate the pre-processing steps of target sentences.



I	school	to	go	presentcont

njAn~	schooL~	lekku	poyi	Kondirikkunnu
nAnu	shAle	ge	hOgu	uttiddEne

Figure 9. Pre-processing step for target sentences.

3) Training the translation system: The well organized pre-processed source and target language sentences were created and used to train the translation system. Training the translation system includes two steps as follow:

a) Creation of Language model: A toolkit called SRILM was used for building and applying various statistical language models. A well organized pre-processed monolingual target language data such as Malayalam or Kannada was given as input to SRILM. The main objective of SRILM is to support language model estimation and evaluation.

b) Creation of Translation model (Phrase table): The translation model was built with another SMT toolkit called GIZA++. In this case a well organized pre-processed bilingual parallel corpus of English and Malayalam (or Kannada) was given as input to the GIZA++. Using an unsupervised learning algorithm called Expectation–Maximization algorithm, GIZA++ computes the translation probabilities from the parallel corpus.

4) Testing the Translation System: Once the translation system was built, the next step is to translate new English sentence into Malayalam or Kannada and find out the performance of the system. The MOSES decoder of SMT was used to translate sentences from source to target language. The following sequence of steps are used for translate and test new English sentences.

Step 1: Pre-process the English sentence and convert it into format suitable to the proposed translation system as explained earlier.

Step 2: The decoder (MOSES) takes this pre-processed English sentence as input. Using the translation model (phrase table) and language model, the decoder decodes the English data to get corresponding Malayalam (or Kannada) sentence.

Step 3: Apply morphological generator on this Malayalam (or Kannada) decoder output to combine the morphemes to form meaningful words of the equivalent target sentence.

Step 4: Using a Roman to Unicode converter convert the target sentence in Unicode text form.

V. Results and Evaluation

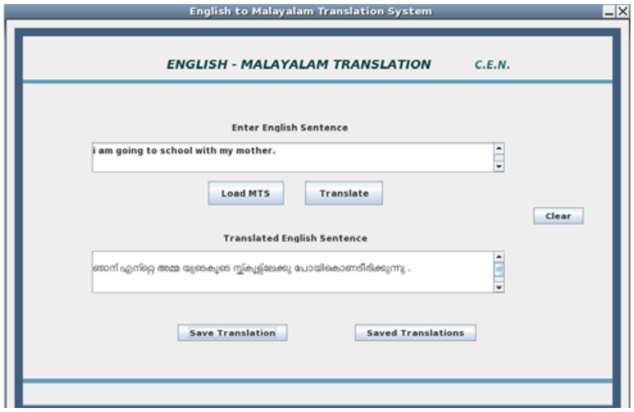
We have created a well organized pre-processed corpus with very simple sentences for training and testing the system. The training and testing sentence size were limited with maximum of twelve words. According to the structure of sentence we have written reordering rules. The statistics of data that we have been used in our translation system is shown in table 10.

TABLE X. Corpus Statistics of Proposed Translation System

Translation system	English to Malayalam		English to Kannada	
Corpus size	Number of Sentence	Number of Words	Number of Sentence	Number of Words
Training	1000	5210	1000	5162
Testing	100	643	100	628

The quality of translation obtained after preprocessing was found to be very promising and extremely high compared to base line translation systems.

The system was able to performed good translation even for simple sentences having more than ten words. From our experiment it found that efficiency of our system with reordering of sentence and adding morphological information have a better performance when compared with a system that uses data without reordering. Fig. 10 shows a sample English to Malayalam translation system screenshot for a sentence 'I am going to school with my mother'.


English to Malayalam Translation System
ENGLISH - MALAYALAM TRANSLATION
C.E.N.
Enter English Sentence
Load MTS
Translate
Clear
Translated English Sentence
Save Translation
Saved Translations

i am going to school with my mother.

ഞാന് എന്റെ അമ്മ യുടെകൂടെ സ്കൂള്ളിലേക്കു പോയികൊണ്ടിരിക്കുന്നു .

Figure 10. Pre-processing step for target sentences.

Also we observed that the proposed translation system reduces the required training corpus size with a greater amount when compared with that system without reordering and morphological information. The performance of proposed translation system was evaluated with BLEU evaluation metric. The table 11 shows a comparison of performance statistics obtained for the proposed translation system with the baseline system.

TABLE XI. Evaluation Statistics of Proposed System

Technique	**BLEU evaluation metric**	
	Blue score for Malayalam	**Blue score for Kannada**
Baseline	15.9	15.4
Baseline + syntax	19.3	19.0
Baseline + syntax + Morphology	24.9	24.5

'Baseline' stands for simple phrase based system; 'Baseline + Syntax' stands for the results after re-ordering and 'Baseline + Syntax + Morphology' stands for the results after morphological processing followed by re-ordering. In all the cases training was performed with the same corpora and the same set of sentences was used testing. The blue scores obtained were low as the training corpus size that we used were very less and obviously depend on the testing data that we used. The performance evaluation shows that when we applied reordering, the blue score increased by approximate 3.5 and when morphological information added that has been again increased by approximate 5.5. In addition to the improved performance the proposed translation system reduces the required corpus size with greater amount.

VI. Conclusion and Future Work

In this Project work we have presented an effective methodology for English to Dravidian language phrase-based statistical machine translation. The results show that significant improvements are possible by incorporating syntactic and morphological information to the corpus. From the experiment we found that the proposed translation system successfully works for almost all simple sentences in their twelve tense forms, their negatives and question forms.

The performance of the proposed system can be improved by adding more transfer rules to cover more classes of sentences. As Dravidian languages like Malayalam and Kannada are very much morphologically rich and agglutinative, the performance can be further improved by adding more morphological inflections to the system. Since all South Dravidian languages have almost the same structure, the approach can be easily extended to Tamil and Telugu. The sentences that we deal with may not be simple always. When the sentences are too long, parser may fail to give the proper syntactic structure. This problem can be solved by developing some intelligent mechanism to split sentences into meaningful small sentences and then solve individually. Since all Indian languages follow SOV order, and are relatively rich in terms of morphology, the methodology presented in general applicable to English to Indian language SMT.

ACKNOWLEDGMENT

We acknowledge our sincere gratitude to Dr. Muralidhar (Professor, Malayalam Department, Govt. Victoria College, Calicut University, Kerala, and India.) and Prof. M Shankaranarayana Bhat (Head of Kannada department and Principal, Junior College, Sampaje, Coorg, Karnataka, India) for their excellent support to generate well organized Malayalam and Kannada corpora. We also express our gratitude to Mr. Shivapratab Gopakumar (Assistant. Professor, CEN, AMRITA Vishwa Vidyapeetham, Coimbatore, India), Mr. Rakesh Peter (Assistant. Professor, CEN, AMRITA Vishwa Vidyapeetham, Coimbatore, India) and Mr. Ajith V P (Research Associate, CEN, AMRITA Vishwa Vidyapeetham, Coimbatore, India) for their valuable support and encouragement for developing these prototype versions of English to South Dravidian languages like Malayalam and Kannada statistical machine translation system.

REFERENCES

[1] Marie-Catherine de Marneffe and Christopher D. Manning, "Stanford typed dependencies manual", nlp.stanford.edu/software/ dependencies_manual.pdf.

[2] Dorr, E Hovy, Los Angelos and L.Levin, "Machine Translation Interlingual Methods", Proceedings of EMNLP, 2000.

[3] Manning and Schutze, "Foundations of Statistical NLP", Proceedings of HLT/NAACL 2003.

[4] Franz Josef Och, Christoph Tillman, and Hermann Ney, "Improved Alignment Models for Statistical Machine Translation", Proceedings of EMNLP, pp:20–28, 1999.

[5] Durgesh Rao, Kavitha Mohanraj, Jayprasad Hegde,Vivek Mehta, and Parag Mahadane, "A Practical Framework for Syntactic Transfer of Compound-Complex Sentences for English-Hindi Machine Translation", Proceedings of KBCS, 2000.

[6] Kenji Yamada and Kevin Knight, "A Syntax-based Statistical Translation Model", Proceedings of ACL, 2001.

[7] Daniel Marcu and William Wong, "A Phrase-base Joint Probability Model for Statistical Machin Translation", Proceedings of EMNLP, 2002.

[8] Philipp Koehn, Franz Joseph Och, and Daniel Marcu "Statistical Phrase-Based Translation", Proceedings of HLT/NAACL 2003.

[9] Franz Josef Och, "Minimum Error Rate Training in Statistical Machine Translation", Proceedings of ACL, 2003.

[10] B. Pang, K. Knight, and D. Marcu, "Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences," NAACL-HLT, 2003.

[11] Kenji Imamura, Hideo Okuma, Eiichiro Sumita, "Practical Approach to Syntax-based Statistical Machine Translation", Proceedings of MT-SUMMIT X, pp:90-95,2004.

[12] I. Dan Melamed, "Statistical Machine Translation by Parsing", Proceedings of ACL, 2004.

[13] Collins, M., Koehn, P, & Kucerova.I, "Clause restructuring for statistical machine translation" ACL, 2005.

[14] Sonja Nieben and Hermann Ney, "Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information", Computational Linguistics, Vol.2, pp:181–204, 2004.

[15] Maja Popovic and Hermann Ney, "Statistical Machine Translation with a Small Amount of Bilingual Training Data", 5th LREC SALTMIL Workshop on Minority Languages, pp: 25–29, 2006.

[16] Michael Collins, Philipp Koehn, and Ivona Kucerova, "Clause Restructuring for Statistical Machine Translation", Proceedings of ACL, pp: 531–540, 2006.

[17] Maja Popovic and Hermann Ney, "POS-based Word Reordering for Statistical Machine Translation", Proceedings of NAACL LREC, 2006.

[18] Marcu, D., Wang, W., Echihabi, A., & Knight, K, "Spmt: Statistical machine translation with syntactified target language phrases", Proceedings of EMNLP, Sydney, Australia, pp:44-52, 2006.

[19] Nizar Habash, "Syntactic Preprocessing for Statistical Machine Translation", Proceedings of the Machine Translation Summit (MT-Summit), 2007.

[20] Ibrahim Badr, Rabih Zbib, and James Glass, "Segmentation for English-to-Arabic Statistical Machine Translation", Proceedings of ACL/HLT, 2008.

[21] Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah and Sasikumar M, "Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation",IJCNLP, 2008.

[22] www.languageinindia.com Vol 6 : 8 August, 2006.

[23] T. N. Vikram & Shalini R Urs, (2007), "Development of Prototype Morphological Analyzer for the South Indian Language of Kannada", Lecture Notes In Computer Science: Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers. Vol. 4822/2007, 109-116.

[24] http\\www.iloveindia.com\languages of India.

[25] K. Narayana Murthy, "Computer Processing of Kannada Language**"**, University of Hyderabad.

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

[27] Philipp Koehn., "MOSES a Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models: User Manual and Code Guide", University of Edinburg, UK, (2009).