10.1 Introduction

Work in the area of Machine Translation in India has been going on for several decades. During the early nineties, advanced research in the field of Artificial Intelligence and Computational Linguistics made a promising development of translation technology. This helped in the development of usable Machine Translation Systems in certain well-defined domains. It is extremely difficult to build a fully automatic high quality machine translation system (FGH_MT). In fact there is no system in the world which qualifies to be called FGH_MT. Many institutions like IIT(Kanpur), CDAC(Mumbai), CDAC(Pune), IIIT(Hyderabad), etc. are engaged in development of MT systems under projects sponsored by Department of Electronics, state governments etc. since 1990. Research on MT systems between various Indian and foreign languages and also between Indian languages are going on in these institutions. Translation between structurally similar languages like Hindi and Punjabi is easier than that between language pairs that have wide structural difference like Hindi and English. Having many parts of their grammars and vocabularies in common, it is easy to develop translation systems between closely related languages.

Development of Machine translation (MT) system requires very close collaboration among linguists, professional translators and computer engineers. In the development process, there are two major goals; (a) accuracy of translation and (b) speed. Accuracy-wise, smart tools for handling transfer grammar and translation standards including equivalent words, expressions,
phrases and styles in the target language are to be developed. The grammar should be optimized with a view to obtaining a single correct parse and hence a single translated output. Innovative use of corpus analysis, efficient parsing algorithm, design of efficient data structure and run-time frequency-based rearrangement of the grammar are required to reduce the parsing and generation time.

Language translation between Malayalam and other languages can be said to be in very primitive stage now. Turning to Machine translation from Malayalam to English, it can be seen that there has been hardly any work done in the area. This was probably the first of its kind with respect to Malayalam language. It was also observed that machine translation systems developed for other language pairs needs improvements in rule set, dictionary design and translation methodology.

The main objective of this research was to study the issues in Malayalam to English translation and the development of a prototype system for Malayalam to English translation. Major tasks of the research work included: i) Study of the morphology of various lexical categories of Malayalam and English ii) Identification of the part of speech tag set and chunk tag set for word sense disambiguation and structure transfer iii) Derivation of syntactic structure of Malayalam and English sentences using morphology and word level dependencies iv) Determination of the syntactic structure differences between Malayalam and English sentences v) Design of efficient lexicon in terms of speed of retrieval, space requirement and speeding up of the translation process vi) Improving speed and accuracy of the translation process by choosing the best models for the different modules for the prototype translator like morphological analysis, parsing etc.
Design and development of the prototype translator

As a first step we conducted a survey on the various approaches tried by various researchers for machine translation between two languages including both Indian and non Indian languages. Researchers used different formalisms that were best suited to their applications. They were mainly i) Direct translation approach ii) Rule based approach and iii) Corpus based approach. Direct translation is appropriate for structurally similar languages. Among the rule based approaches transfer based systems based on syntactic structure transfer are more flexible and it can be easily extended to language pairs in a multilingual environment. The interlingua based systems can be used for multilingual translation. Recently, the Universal Networking Language has been proposed as the interlingua by the United Nations University for overcoming the language barrier. Over the past decade data-driven approaches to machine translation have widely been used in language processing research. The relative success in terms of robustness of Example Based and Statistical approaches have given rise to a new optimism and an exploration of other data-driven approaches such as Maximum Entropy language modeling. Performance of statistical techniques can be improved through large parallel corpus and usage of linguistic knowledge in the model. Hybrid systems are found to have better performance compared to the ones with the component technology.

It was found that there were only very few systems developed for south Indian languages. More research has to be done in these areas to overcome the language barrier faced by India. The MT systems developed had many shortcomings in terms of rule set, dictionary, translation methodology and it was apparent from the survey that further work is needed in MT as a whole to produce intelligible translations. From among the various approaches the transfer based approach
has been chosen due to its flexibility and extensibility to other language pairs in a multilingual environment.

The following tasks were performed for the design of the various modules of the prototype systems like morphological analyser, parser etc. and for the development of the translator as a whole.

a) Study of the morphology of various lexical categories of Malayalam and English:

The morphological variations for words in Malayalam were studied. Malayalam belongs to the category of highly agglutinative languages in which formation of new words by combining a noun and a noun, noun and adjective, verb and noun, adverb and verb, adjective and noun, and in some cases all the words of an entire sentence to reflect the semantics of the sentence are very common. In addition to this nouns and verbs show inflectional morphology. The nouns are inflected because of gender, number and case. Verbs show inflections due to tense, aspect and mood information. Derivational morphology is also found for verbs and nouns in which other lexical categories like adjectives, adverbs and infinitives etc. are formed from verbs and nouns by adding proper suffixes. The prototype system makes use of most of the commonly found rules for inflection, derivation and word compounding for the purpose of reducing the dictionary space used.

b) Identification of the part of speech tag set and chunk tag set for word sense disambiguation and structure transfer.

POS tagging is an important activity for investigators of natural language processing, speech recognition and other related areas. It proves to be a basic building block for constructing statistical models which needs an annotated corpora for automatic processing of natural languages. The words in a sentence
occur as groups called chunks. In problems like machine translation or language understanding these chunks play a very important role. We used IIIT tag set as a benchmark. A set of primitive morpheme tags and chunk tags required for the translator were derived.

c) Derivation of syntactic structure of Malayalam and English sentences using morphology and word level dependencies.

According to Universal clause structure grammar clauses in a sentence can be nested one inside the other, resulting in a hierarchical or tree like structure. Verb groups and sentinels contain all the required information for recognizing clauses, for determining the nested or hierarchical structure of clauses and for determining the clause boundaries. It is seen that every clause in a sentence except for the main clause has a sentinel which marks one of the boundaries of that clause. A careful analysis of the different sentence classes of Malayalam and English was required for deriving the hierarchical dependency rules and also the syntactic structure transfer rules of the prototype translator. Malayalam is an S-O-V language. The default or unmarked order of constituents is Subject first, then the Object and finally the verb. However, Malayalam, being a relatively free word order language, permits substantial amount of freedom in the order of constituents although normally the verb remains in the sentence final position. Word order becomes less important mainly because noun groups are marked for cases and the verb agrees with the subject in gender, number and person. In fact, subjects and objects are often dropped. Normally all modifiers precede the modified. There are a variety of subordinate clauses. Subordinate clauses also precede the main clause. They typically involve special non-finite forms of verbs which occur invariably in the clause final position and mark the
right hand boundary of the respective clauses. All these assertions were taken as rules.

d) Determination of the syntactic structure differences between Malayalam and English sentences.

13 rules were identified and implemented in the prototype system. These rules cover most of the commonly found structure difference between Malayalam and English.

e) Selection of a suitable model for the translator.

The survey on the various approaches for machine translation revealed that a direct translation approach and corpus based approaches will be inefficient in the case of Malayalam to English translation system. So we have decided to use a syntax transfer based approach using synchronous tree adjoining grammar (STAG). Translation with Synchronous TAGs is a three step process:

1. Derivation tree for source sentence is obtained by parsing the source sentence.

2. Source derivation tree is converted into target derivation tree or trees.

3. The target sentence is obtained by listing the leaf nodes in the target parse tree in a depth first order.

The system modules selected are like that of SAMPARK system developed by IIIT Hyderabad. The syntax structure formalism used is similar to the one used in UCSG based Kannada translator by Murthy. The parsing method is similar to the one in MANTRA system developed by CDAC Mumbai. We had tried to
combine modern artificial intelligence techniques with the classical Paninian framework based on Sanskrit grammar.

There are four main modules for the system: i) Morphological analyzer ii) Morpheme based parser iii) Target sentence generator and iv) Bilingual dictionary.

i) **Morphological analyzer**: It finds the sequences of morphemes in the input sentence. Since there are more than one way of splitting a word into morphemes it performs a depth first approach to find all possible sequence of morphemes for the given sentence. The morphological analyser takes a valid sentence of the source language as input and it produces a sequence of morphemes in the sentence as output. It was noted that a two level finite automata was not enough for highly agglutinative language like Malayalam. So a recursive depth first search with backtracking algorithm has been utilized to find all the sequences of morphemes for the given sentence. The algorithm uses rules for morphophonemic changes at each morpheme boundary to find the constituent morphemes. Unlike the previously developed morphological analysers for Malayalam, the morphological analyzer of our system finds only the morpheme sequences in the given sentence. A morpheme can have more than one lexical category and the appropriate category for each morpheme is found only during the parsing stage.

ii) **Morpheme based parser**: It performs three functions. i) It finds the correct POS tag for each morpheme for word sense disambiguation. ii) It identifies the chunks in the sentence for reordering and creates the source parse tree. iii) It performs the reordering required to meet syntactic requirements of the target
language and to create the target parse tree. The parser model selected was a top
down parser based on Synchronous TAG. The model keeps both source and
target tree pair, and performs operations simultaneously while traversing
through the tree nodes. Thus the syntax structure transfer from Malayalam to
English is integrated into the parser. The syntax rules were given in the regular
expression form. The longest rules are listed first on the right hand side of each
production rule so that longer chunks are recognized first. Regular expression
notation helps for a compact representation of language syntax and also it
provides easy modification of the syntax rules.

**iii) Target sentence generator:** The function of the target generation module is to
generate the target sentence by performing a depth first traversal of the target
parse tree created by the parser. It also performs surface form generation of
English words using the morpheme sequences in the tree. It also uses additional
set of rules for the generation of target language sentences in the correct form.

**iv) Bilingual dictionary:** The dictionary is the most essential part of the whole
machine translation system. The space required for the dictionary was
minimized by storing morphemes. The bilingual dictionary contains the
following information:

i) Most of the commonly occurring verbs, nouns, pronouns, adjectives,
inflectional and derivational suffixes, clause suffixes etc.

ii) Each entry in the file has four fields: the root word in Malayalam
(morpheme), the lexical category of the morpheme, category
(human/nonhuman) information (for nouns only) and its translation. A
root word can have more than one tag.
iii) The verbs in past tense have their root words stored along with them.

10.3 Performance Evaluation

The performance evaluation was conducted in three levels: In the analyzer level the parser level and the MT system as a whole. The evaluation was conducted with different classes of texts like news articles, stories etc. The analyser generated three categories of output namely unique correct sequence, multiple sequences with incorrect sequence and multiple correct sequences. The parser also generated three categories of output. The evaluation on the prototype system revealed that the system gives correct answers for most of the sentences belonging to the derived syntax. It generated three kinds of translation: unique correct translation, multiple translations with incorrect translations and incorrect translation. The incorrect translation was due to the limitation of the analyser, the parser or the target sentence generator. The parser needs more syntax rules and more semantic level analysis is required for achieving better results.

10.4 Future enhancements

1. Currently the morphologic analser and parser are two separate modules. The morphological analyser can be integrated into parser thus reducing the number of wrong sequences of morphemes generated by the analyser module.
2. The rules for morphology, the lexical categories for words, the syntax rules and the structure transfer rules are to be enhanced for the system to handle texts belonging to all domains.
3. The parser could perform word sense disambiguation based on lexical category only. To handle other types of sense disambiguation more context dependant rules or statistical approaches are to be incorporated into our system.
4. The system has been implemented in such a way that the parser finds all parses valid for the current input and the translator creates the translations corresponding to each of these parses. The selection of the correct translation is left to the user. This has been one of the drawbacks of the system which should be rectified using appropriate techniques. Use of statistical approaches for parsing with a corpus of correctly parsed sentences can reduce multiple parses greatly. Hidden Markov models have been used for the purpose[17]. The number of wrong parses can be eliminated through the use of unambiguous grammar rules. Preprocessing of the morpheme sequence to eliminate wrong tags for a word examining adjacent tags also can reduce the wrong parses greatly. The sentences with blanks to mark the boundaries of semantically related units also help to reduce wrong parses.

5. Also the bilingual dictionary has to be enhanced by including more words.

10.5 Conclusion

The objective of the research work was to conduct a study on the various issues in a syntax based translation method and selection of appropriate models for the various modules for the system and also to develop a prototype Malayalam to English translator based on the selected models. Encouraging results were obtained even with a small set of resources. Weeding out the limitations of the present work will help to shape the prototype into a full fledged system for machine translation. The method adopted was based on artificial intelligence techniques and can easily be extended to build MT systems for other language pairs.