Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation

Jisha P.Jayan
Sree Sankaracharya University of Sanskrit, Kalady

Rajeev R R
Indian Institute of Information Technology and Management - Kerala

Dr. S Rajendran
Tamil University
Thanjavur

ABSTRACT
Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself. Some prefer the term Computational Linguistics in order to capture this latter function, but NLP is a term that links back into the history of Artificial Intelligence (AI), the general study of cognitive function by computational processes, normally with an emphasis on the role of knowledge representations, that is to say the need for representations of our knowledge of the world in order to understand human language with computers. A morphological analyzer or generator supplies information concerning morphosyntactic properties of the words it analyses or constructs. Morphological Analysis and Generation are important components for building computational grammars as well as Machine Translation. Morphological Analyzer is a program for analyzing the morphology of an input word; the analyzer reads the inflected surface form of each word in a text and provides its lexical form while Generation is the inverse process. Both Analysis and Generation make use of lexicon.

Malayalam like the other languages in the Dravidian family exhibits the characteristics of an agglutinative language. Here using a bilingual dictionary, the Malayalam morphological analyzer and the Tamil morphological generator have been described.

Keywords
Natural Language Processing, Morphological Analyser, Morphological Generator, Malayalam, Bilingual Dictionary, Unicode.

1. INTRODUCTION
The aim of NLP is studying problems in the automatic generation and understanding of natural languages. Natural language is understood as a tool that people use to express themselves, has specific properties that reduce the efficiency of textual information retrieval systems. These properties are linguistic variation and ambiguity. Natural language processing (NLP) is a subfield of artificial intelligence and linguistics. It studies the problems of automated generation and understanding of natural human languages. Natural language generation systems convert information from computer databases into normal-sounding human language, and natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate. Machine translation is a very important application of Natural Language Processing (NLP). Machine translation is throwing up many challenges and opening up many opportunities for doing work. Some of the problems relate to grammars; others pertain to word analysis, bilingual dictionaries, language generation, etc. In machine translation morphological analysis is the main process. First step to analysis the source language. The language faculty in human being has the ability to analyze a given language.

There are various methods by which a morphological analyzer can be built and we propose the Suffix Stripping Method which is found to be very economical. An analyzer can analyze the inflected form of a word into suffixes and stem even if the stem is not entered in the dictionary. The general format of the morphological analyzer of Malayalam is

$$Word \rightarrow stem/root + suffixes$$

The basic principle of morphological generation is to get forms from a root and a set of properties (lexical category and morphological properties). A morphological generator needs to be designed to tackle the different syntactic categories such as nouns, verbs, adjectives, adverbs etc. separately, since the addition of morphological constituents to each of these syntactic categories depends on different types of information. The Suffix Joining Method is used for building morphological generators. The identified suffixes are used along with the morphophonemic rules and morphotactics for developing the morphological generator. The general format of the morphological generator is

$$Stem/root + suffixes \rightarrow Word.$$

2. MALAYALAM AND TAMIL LANGUAGE
The Dravidian Language Family is one of the important groups of languages that are spoken by in South India. There are four recognized Dravidian languages of Telugu, Malayalam, Kannada and Tamil. The most characteristics feature of the Dravidian languages is that they are agglutinative and exhibit the inclusive and exclusive feature.

2.1 Malayalam Language
Malayalam belongs to the Dravidian family of languages and is a language registering a heavy amount of agglutination. The origin of Malayalam as a distinct language may be traced to the last quarter of 9th Century A.D. (Ramaswamy Iyer., 1936). Throughout its gradual evolution the most important influence on Malayalam has been that of Sanskrit and Prakrit brought into Kerala by Brahmins. In modern Malayalam also a good part of vocabulary is of Sanskritic origin. Influence of Sanskrit, the Indo-Aryan language, is evident in the alphabet, phonology and vocabulary and to a lesser extent in morphology also. This dynamic synthesis of diversities has been achieved by no other Indian languages. Malayalam has a special place in the classification of world languages. It is

from Tamil that Malayalam was born. There are different spoken forms in Malayalam even though the literary dialect through out Kerala is almost uniform. Malayalam has its own distinct script, a syllabic alphabet consisting of independent consonant and vowel graphemes plus diacritics.

2.2 Tamil Language

Tamil is the most highly cultivated of the Dravidian tongues, spoken by about fifteen million people in South-East India and Sri Lanka. The origins of Tamil, like the other Dravidian languages but unlike most of the other established literary languages of India, are independent of Sanskrit. The language has a rich and varied vocabulary. It is extraordinary in its subtlety and sense of logic, and the refinements of its grammar are most precise for expressing nuances of thought and meaning. It is remarkably rich in honorific, a characteristic of a decadent rather than a primitive culture.

3. MORPHOLOGICAL ANALYSER

Morphological analysis is the segmentation of words into their component morphemes and (usually) the assignment of grammatical information to grammatical categories and the assignment of the lexical information to a particular lexeme or lemma. Morphological analysis consists of the identification of parts of the words, or more technically, constituents of the words. The design and implementation of morphological analyzer and generator for Malayalam is a promising research for various applications in NLP. Malayalam language is a inflectionally rich in morphology [1], by adding suffixes with the root / stem word.Different methods have been evolved for the implementation of a morph analyzer. Some of the methods used for the analysis of agglutinative languages are Hankamer's Keci [2][Turkish], PC_Kimmo[2] [Finnish], Ample [4] [Quechua]. PC_Kimmo and Ample adopted the root driven approach in the analysis. In root driven method, root/stem is identified at first and the affixes are passed. In affix Stripping method the process takes place in the reverse direction [5]. The affixes are identified first and the remaining part is assumed as the stem or root.

Suffix Stripping [6] which is a method used for analysis makes use of a stem dictionary (for identifying a valid stem), a suffix dictionary containing all possible suffixes that nouns/verbs in the language can have (to identify a valid suffix), morphotactic rules and morphophonemic rules or sandhi rules. This method is economical. Even if the item does not exist in the dictionary, the analyzer can identify the suffixes and stem. Once the suffixes are identified, removing the suffixes and applying proper sandhi rules can obtain the stem. Suffix stripping algorithms do not rely on a lookup tables; instead, rules are stored to find the root/stem.

In highly agglutinative languages such as Malayalam, a word is formed by adding suffixes to the root or stem. Absolutely no prefixes and circumfixes are there in Malayalam. But morphologically highly complex words exist in such languages, which are formed by continuously adding suffixes to the stem. Suffix Stripping method make use of this property of the language, i.e., having complex suffixes attached to the stem. Once the suffix is identified, the stem of the whole word can be obtained by removing the suffix and applying proper sandhi rules.

Malayalam has a tendency to join two words, this is one of the major issue faced. How ever for morphological analysis we need the word split into two. A separate sandhi splitter is necessary to split the combined word in to two. At present there is no sandhi splitter programme. A sandhi splitter demands a morphological analyzer and a morphological analyzer demands a sandhi splitter. There is a dead lock between the two. Availability of morphological paradigms and calcification is another major issue in developing Malayalam Morphological Analyzer. Morphology plays an important role in the overall structure of language. Even most of the syntactic information is embedded as morphological structures. This causes the analysis of word forms of Malayalam to cross the limits of morphology and it reaches to syntactic and semantic level. The written language has its own grammar, which usually differs from the spoken language. Many usages have creped into the written language which violates that grammar. There are many variant spellings for the same word. Some of the spellings reflect dialectal variation; the others however are there simply because of lack of standard conventions.

As Malayalam is morphologically rich and agglutinative with complex structures and there are so many morphophonemic changes in the word formation process that the Root driven and Brute force methods are not sufficient to analyzer/generate words and their forms.

4. BILINGUAL DICTIONARY

Bilingual dictionary is a crucial part not only for machine translation [7], but also for other natural language processing applications such as cross-language information retrieval [8]. Creating a bilingual dictionary in the form of lexemes or words is a difficult task as it covers more than one area of meaning, but these multiple meanings don't correspond to a single word in the target language.

Basically machine translation systems are linked to electronic dictionaries. The content of the dictionaries must be adequate in both quantity and quality: that is, the vocabulary coverage must be extensive and appropriately selected [9], and the translation equivalents carefully chosen [10], if target language output is to be satisfactory or indeed even possible. The size and quality of dictionary limits the scope and coverage of a system, and the quality of translation that can be expected [11]. The dictionary entries are based on dictionary entries for lexical stems of specified category, strictly monolingual analysis and generation dictionaries, and transfer dictionaries based on language-pair-specific information.

5. MORPHOLOGICAL GENERATION

The aim in morphological generation is to produce the inflected form of a word according to the features and values in the Feature Structure. It is also necessary to reuse the linguistic resources created for analysis purpose. From practical point of view, morphological generation is the inverse process of analysis, namely the process of converting the internal representation of a word to its surface form. The same rule definitions can be used to generate the desired word form as used for analysis. The only difference will be the direction of execution order of the elements in the rule definition. For example:

Root: MOUSE

category (PartOfSpeech): Noun Number: Plural

Stem: MOVE

category (PartOfSpeech): Verb Tense: Past

If our internal representations of the words *mice* and *move* are as shown above, then morphological generation would convert these to the character strings *mice* and *moved.*

The morphological generation mainly deals with the concatenation of corresponding suffixes with the root word to form a word of specific grammatical category. The input of the morphological generator would be the root word which then inflects this word to the morphology of the respective language and gives as the output the target forms of the word. The Morphological structure of Tamil verb is quite complex since it caters to person, gender, and number markings and also combines with auxiliaries that indicate aspect, mood etc While morphologically generating the verb, the gender, number and person of the subject is necessary in order to select the appropriate suffix catering to the selected tense. So while going from Malayalam to Tamil, there are about eleven different forms for a single stem in Tamil

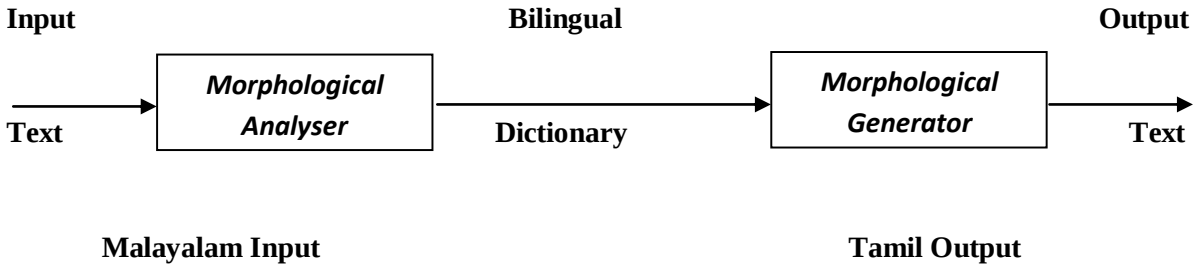

Fig 1: Block Diagram of Malayalam – Tamil Word Translation

6. IMPLEMENTATION

The bilingual dictionary for Malayalam and Tamil consist of the root/ stem of the words with its grammatical category. The suffix stripping method has been used for developing the morphological analyser and for developing the morphological generator, the suffix joinning method has been used. The program has been done using the PERL script. The input given to the analyser and the output obtained from the generator are in Unicode notation.

6.1 Algorithm

Step 1: Get the word to be analyzed.

Step 2: Check whether the entered word is found in the Root Dictionary.

Step 3: If the word is found in the dictionary, stop;

Else

Step 4: Separate any suffix from the right hand side

Step 5: If any suffix is present in the word, then check the availability of the suffix in the dictionary.

Then

Step 6: Remove the suffix present, then re-initialize the word without identified suffix ,

Go to Step 2.

Step 7: Repeat this process until the Dictionary finds the root/stem word.

Step 8: Store the Malayalam root/stem word in a variable and then get the corresponding Tamil word from the bilingual dictionary

Step 9: Check what all grammatical features does the Malayalam word have given and then generate the corresponding features for the Tamil word

Step 10: Exit.

6.2 Result:

The entered word is വരും (varuM)

[Stem] വര് (var)

[Future Tense] ഉം (uM)

Tamil Equivalent for the Given Malayalam Word is :வா (vaa)

The Generated Tamil Words are Following

வருவோம் (varuvoom)

வருவாய் (varuvaay)

வருவீர் (varuviir)

வருவீர்கள் (varuviirkaL)

வருவான் (varuvaan)

வருவாள் (varuvaaL)

வருவார் (varuvaar)

வருவார்கள் (varuvaarkaL)

வருவான் (varum)

7. CONCLUSION

A significant part of the development of any machine translation (MT) system is the creation of lexical resources that the system will use. Dictionaries are of critical importance in machine translation. A bilingual dictionary or translation dictionary is a specialized dictionary used to translate words or phrases from one language to another. They are the largest components of an MT system in terms of the amount of information they hold. The proper functioning of a morphological generator necessitates efficiency in the generation of a word, once provided its Root or Stem and the corresponding feature values. The Suffix Stripping for morphological analyzer and the Suffix Joining for the morphological generation of words proved to be an efficient method. Since words are formed by the suffix addition with root, most of the words can take the POS tag based on the root or stem. Hence in Malayalam the suffixes play major role in deciding the POS of the word.

8. ACKNOWLEDGMENTS

We would like to thank Ministry of Communication and Information Technology, Department of Information Technology, Government of India, for promoting a project on Indian Language to Indian Language Machine Translation System, where this part of the work was carried out. Large part of this work was carried out at Tamil University, and our thanks goes to all staff of ILMT Lab, Department of Linguistics who gave us all guidance for this developing this.

9. REFERENCES

[1] Sumam Mary Idicula and Peter S David, A Morphological processor for Malayalam Language, South Asia Research, SAGE Publications, 2007.

[2] J. Hankamer , Finite state morphology and left to right phonology, Proceedings of the Fifth West Coast Conference on Formal Linguistics, Stanford, CA, 1986, pp 29-34.

[3] E. Antworth. PC-KIMMO: A two-level processor for morphological analysis ,Dallas, TX: Summer Institute of Linguistics.1990

[4] D.J. Weber, H.A. Black & S.R. McConnel, AMPLE: a tool for exploring morphology, Dallas, TX: Summer Institute of Linguistics.1988

[5] Gülsen Eryioit and Esref Adalý. Proceedings of International Conference on AI and Applications, Austria. 2004. pp 299-304.

[6] Rajeev R,R, Rajendran N and Elizabeth Sherly. A Suffix Stripping Based Morph Analyser For Malayalam Language, Science Congress 2008 pp 482-484.

[7] F.Och and H.Ney. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics. 2003. pp 29(1): 19-51.

[8] G.Grefenstette. The Problem of Cross-language Information Retrieval. Cross-language Information Retrieval. Kluwer Academic Publishers. 1998. pp 1-9.

[9] Ritchie, Graeme. The Lexicon. In Whitelock, eds.1985.p. 225-256.

[10] Knowles, Francis. The Pivotal Role of the Dictionaries in a Machine Translation System. In Lawson, Veronica, ed. Practical Experience of Machine Translation. North-Holland. 1982.

[11] D. Arnold, L. Balkan, S. Meijer, R. L. Humphreys, and L. Sadler. Machine Translation: An Introductory Guide, ch.5. NCC Blackwell, UK 1994.