DIFFICULTIES IN PROCESSING MALAYALAM VERBS FOR STATISTICAL MACHINE TRANSLATION

Jayan V[1], Bhadran V K[2]

[1&2]Language Technology Centre, Centre for Development of Advanced Computing(C-DAC), Thiruvananthapuram, Kerala, India

ABSTRACT

In this paper we discuss the difficulties in processing the Malayalam texts for Statistical Machine Translation (SMT), especially the verb forms. Mostly the agglutinative nature of Malayalam is the main issue with the processing of text. We mainly focus on the verbs and its contribution in adding the difficulty in processing. The verb plays a crucial role in defining the sentence structure. We illustrate the issues with the existing google translation system and the trained MOSES system using limited set of English-Malayalam parallel corpus. Our reference for analysis is English-Malayalam language pair.

KEYWORDS

Causatives, postposition, verb morphology, morphemes

1. INTRODUCTION

Malayalam is a language spoken in India which is spoken in the state of Kerala. It belongs to Dravidian language family and is spoken by more than 38 million people. Malayalam is highly agglutinative in nature. It is roughly estimated that a whopping eighty percentage of the vocabulary of the scholarly usage of the languages like Malayalam is constituted by Sanskrit[1]. Tamil and English also influenced Malayalam in one or the other. Inflection, derivation, compounding and concatenation are the major morphological behaviours in Malayalam. According to grammar rules, Malayalam words are divided into two main types; *vachaka* and *dyothaka*, *dyothaka* denotes relation and it has no individual meaning. *Vachaka* is split to three types noun, verb and qualifying words. The first step of natural language processing(NLP) is to recognize the words in a sentence. We have to consider the way they are created, placement of morphemes in a word, combinations of the morphemes or words and the rules associated with the formation of a semantic category. The analysis of words will provide the complete syntactic and semantic information. In natural language processing Malayalam is still in the nascent stage because this piece of technology is not much popular among the scholars and computing people. The main reason behind this is lack of available resources. There must be an accord between the language rules and computation. For the analysis based on statistics we need a huge parallel corpus that can be processed readily. The first step of natural language processing is to recognize the words in a sentence. We have to look in to the way they are created, placement of morphemes in a word, combinations of the morphemes or words and the rules associated with the formation of a semantic category. The analysis of words will provide the syntactic and semantic information. There must be an accord between the language rules and computation. NLP requires contributions from both language scholars and computer specialists. An effective linguistic rule customized for computational purpose can make a paradigm shift in the field of NLP.

Verbs are the main factor which contributed to today's language form. Verbs normally reflect the strength and ability of a language. The interaction between languages results in acquiring the nouns and adjectives from the other language. Here adopting of verbs are very rare. But the interaction of Malayalam with Sanskrit resulted in borrowing verbs also in large extent. Suranad Kunjan Pillai identified 2880 verbs for the classification. In that around one third of the verbs are the loan words from Sanskrit. In addition to that in Malayalam verb forms also formed by the combination of Sanskrit noun and the verb forms like '*petuka*', *ceyyuka*,etc. Out of the 1013 sanskrit verbs only about 300 roots are taken from Sanskrit. Rest of them are formed by affixing suffixes and prefixes along with this root forms.[2] In addition to this 2880 verbs are also formed by combining the verbs. For example for the verb 'to discover' the Malayalam meaning is '*kaṇṭupiticcu*'. It is a combination of verbs '*kaṇṭu*'(see+past) and '*piṭiccu*'(catch+past). When they combined together, that will have a different meaning as 'to discover'. Likewise Malayalam has many compound verbs formed by combining some of the 2880 root words. This will add to the complexity of Malayalam morphological Analyzer.

2. RELATED WORKS

Most of the works carried out based on the analysis of the parallel corpus using the MOSES system. Most of the SMT systems are using the phrase based models as their translation approach and mainly depends on the mapping of chunks without considering any linguistic information. MOSES system takes care of the linguistic information to some extent. A study on the out of vocabulary word handling is mentioned in [3]. This will find the out of vocabulary (OOV) word after going through the phrase table and handle this word from the OOV handler module. Then it will pass on to Translation model with extended phrase table and then go for decoding. Another paper [4] focussed on improvement of word alignment. Techniques to improve the word to word alignments between the English - Malayalam sentence pairs are discussed in this paper. Using the parts of speech tags as an additional knowledge source, the parallel corpus is enriched to contain more information for selecting the correct translation for a Malayalam word. The alignment model with category tags is useful in diminishing the set of alignments for each sentence pair and thereby simplifying the complexity of the training phase. The named entities and cognates located in the sentence pairs also have an important role in reducing the insignificant alignments. But in many cases only the data sparsity due to the suffixes in nouns are discussed. That can be readily handled using the factored model of the MOSES system. But we could not find any study or results based on the verb influence on the SMT system.

3. VERB ANALYSIS FOR SMT

Verbs play a crucial role in identifying the tense, aspect and modality of a sentence. The interrogation, negation and conditional markers are also attached as suffix in the verb form. Malayalam verbs do not exhibit concord with noun phrases that are their arguments. The conjunction marker is also suffixed to the verb when multiple verbs are coming in a sentence. So the analysis of verb is inevitable in any kind of Machine Translation (MT) system. Now we are introducing different cases where the MT system needs to be concentrated while translating the Malayalam sentences and vice-versa. It will be applicable for all kinds of MT systems viz. Rule based, example based and SMT. In Malayalam, based on Suranad Kunjan Pillai's verb classification there are 16 verb classes. He classified the verbs based on their past form. There are 12 classes of verbs ending with '*tu*' and 4 classes of verbs ending with '*i*'. Morphophonemic changes are occurring while the root word is combining with the past form of the verb. For computational purpose these verb classes further classified in to 52 classes by Jayan et.al[4].

3.1 Verb Inflections based on Tense

Table 1. Tense forms of verbs

Root Word	Past	Present	Future
എഴുത്(*eḻut*)	എഴുതി(*eḻuti*)	എഴുതുന്നു(*eḻutunnu*)	എഴുതും(*eḻutuṁ*)
കൺ(*kaṇ*)	കണ്ടു(*kaṇtu*)	കാണുന്നു(*kāṇunnu*)	കാണും(*kāṇuṁ*)

In the table we can see that the two forms of verbs that are ending with 'i' and 'tu' in the past tense form. Due to some morphophonemic changes occurring while sandhi formation, in the second case we can find a different form. The sandhi process will be as shown below:

എഴുത്+ ഇ → എഴുതി
eḻut + i → eḻuti
(Write+PAST) (wrote)
കൺ+തു → കണ്ടു
kaṇ +tu → kaṇtu
(See+PAST) saw

3.2 Verb Inflections based on Aspect

Now in the table below we are focusing on the verb '*eḻut*'. The list of forms are listed below along with their tense forms

Table 2. Aspect forms of verbs

Aspect	Past	Present	Future
Continuous	*eḻutukayāyirunnu* *write+INFIN-PAST* was writing	*eḻutukayākunnu* *write+INFIN+PRES* is writing	*eḻutikkoṇtirikkuṁ* *write+PROG+FUT* shall be writing
Perfect	*eḻutiyittuṇtāyirunnu* *write+PERF+PAST* had written	*eḻutiyittuṇt* *write+PERF+PRES* has written	*eḻutiyittuṇtākuṁ* *write+PERF+FUT* will have written
Perfect Continuous	*eḻutikkoṇtirikkunnuṇtāyirunnu* *write+PROG+PERF+be + PAST* had been writing	*eḻutikkoṇtirikkunnuṇt* *write+PROG+PERF+be + PRES* has been writing	*eḻutikkoṇtirikkunnuṇtākuṁ* *write+PROG+PERF+be+ FUT* will have been writing

Above table clearly indicates the complexity of the Malayalam verbs in one form. We have different forms of aspect viz. perfect and Imperfect. Again in perfect aspect we have three forms of aspects in Malayalam. This will make the verb much more complex in the analysis part.

3.3 Verb Inflections based on Mood

Verb forms in conditional clauses are formed by the addition of the suffix *–āl* to the past tense stem or by the addition of *-eṅkil* to any of the three tense forms. For the wish for something to happen, the suffix *-aṭṭe* is added to the root word. Similarly obligation is expresses by *-aṇaṁ* and

the negative *-aṇṭa* are added to the root verb. Other forms of moods are *–arut*, infinitive+*pātilla,*etc. All these forms are attached with the verb.

For example:

(1) ini ninakk pōkāṁ
Now you go-PERMIS
Now you may go

There are many such forms of mood that is attached to the verb also attributing to the complexity of verb.

3.4 Causativization of Verbs

Inherently verb may be transitive or intransitive. The valency of either set can be increased by changes in the syntactic structure of the sentence or by modification in the verb stem or by the combination of both of these. The morphological change takes the form of (i) modification of the final consonant of the verb root, or (ii) the addition of a causative suffix [8].
The forms which give causativisation are as follows:

X opens
J causes X to open
K makes J cause X to open
L makes K make J cause X to open

Consider the examples:

(2) tala tuvarnnu
head become dry-PRES
The head becomes dry.
(3) ñān enṟe tala tuvartti
I I-ACC head dry-TRANS-PAST
I dried my head
(4) amma kuññinṟe tala tuvartticcu
mother child-ACC head dry-TRANS-CAUS-PAST
Mother dried the child's head.
(5) amma vēlakkāriye koṇt kuññinṟe tala tuvarwwippiccu
mother servant-INSTR child-GEN head dry-TRANS-CAUS2-CAUS1-PAST
Mother got the servant to dry the child's head

The causative marker '*iccu*' and the double causative marker '*ippiccu*' determines the suffixes to be attached in the objects of the sentence. In case (3), '*nṟe*' is the suffix with accusative case. In the case (4), we can see that there are two objects and for the first object a case marker along with a postposition is added and for the second object case marker is added. So the causative marker attached to the verb decides the suffixes to be attached in the objects. But in an SMT with phrasal chunks as the main fuel for constructing the sentence the case marker selection will not be feasible.

The morphology of transitivisation and causativisation in Malayalam is somewhat under researched subject.

3.5 Negation attachment with verb

The negation of the sentence can be realized using the verb. So the translation becomes difficult using SMT for sentence with negation marker. Inserting a particular negative word in the target language becomes difficult. So we need some alternate solution for handling the negation in SMT. Consider the sentence below:

(6) *rāvile enikk oru kapp cāya matiyāvilla*
morning I-DAT a cup tea insufficient
In the morning a cup of tea is not enough for me.

(7) *at oru nalla kaḷi āyirunnilla*
that a good game be-NEG
That was not a good game.

(8) *āruṁ paripātiyil paṅketuttilla*
nobody program-LOC participate-NEG
Nobody participated in the program

In sentence (8) we can see that the negative sense in the word is reflected in the verb. This is the real problem in the word or phrase alignment in SMT system.

3.6 Verbs in conjunction

When two or more verbs are there in a sentence in conjunction, then that will be another major issue with SMT system. Let us look in to the example below:

(9) Ram played and had food.
rāman kaḷikkukayuṁ bakṣaṇaṁ kaḷikkukayuṁ ceytu
Ram play-PAST-CONJ food eat-PAST-CONJ do-PAST

In the sentence above we can find that conjunction marker is attached with each verb and an additional verb, '*cey*' is introduced in to the sentence. This will really affect the translation in both ways.

4 GOOGLE TRANSLATE AND ITS OUTCOMES

Google Translate is a multilingual service provided by Google Inc. to translate written text from one language into another. It supports 90 languages. Google Translate does not apply grammatical rules, since its algorithms are based on statistical analysis rather than traditional rule-based analysis. The system's original creator, Franz Josef Och, has criticized the effectiveness of rule-based algorithms in favour of statistical approaches The above mentioned cases were tried with the Google Translate SMT. It is based on a method called statistical machine translation.
Google Translate has its limitations like other automatic translation tools. The service limits the number of paragraphs and the range of technical terms that can be translated, and while it can help the reader to understand the general content of a foreign language text, it does not always deliver accurate translations. Grammatically, for example is Google Translate fails to differentiate between *imperfect* and *perfect* tenses in Romance languages so habitual and continuous acts in the past often become single *historical* events. Knowledge of the *subjunctive mood* is virtually non-existent.

The system may be working for simple tense forms. But when it comes to the aspects the system fails to translate. For example consider a sentence for simple present tense.

(10) *avan eḻutunnu*
he write-PRES
he writes

Similarly if we go for present perfect continuous we can see that the system fails to translate and we will get the source sentence at the output. This system has also problem with the modals. Consider the example below:

(11) Now you may go
Google output
ippōḷ ningal pōkuvān
now you go-INF
Actual Output
ippōḷ ningal pōyāluṁ
now you go-MOD

Considering causativation the Google system also fails incorporate all the features to be added in the verb root. Consider the example below:

(12) I dried my head
Google output
ñān enṟe tala uṇaṅṅiyirikkunnu
I I-ACC head dry-PERF-PAST
Actual Output
ñān enṟe tala tuvartti
I I-ACC head dry-TRANS-PAST

(13) Mother dried the child's head.
Google output
amma kuttiyute tala uṇakkunnu
mother child-ACC head dry-PRES
Actual Output
amma kuttiyute tala tuvartti
mother child-ACC head dry-CAUS-PAST

We can see that in google the system generates the simple present verb output instead of the causative past. This will be a real challenge to incorporate in the SMT system. When it comes to the double causative sentences, the system fails completely and gives completely wrong translation. It is mainly due to the causative suffix attachment to the verb stem.

Now consider the sentences with negation and the conjunction of verbs.

(14) Nobody attended the program
Google output
āruṁ paripāti paṅketuttu
nobody program attend-PAST
Actual Output
āruṁ paripātiyil paṅketuttilla
nobody program-LOC attend-NEG-PAST

Here we can see that the verb is not taken the negation in the verb. It is also another issue that is to be handled.

(15) Raman drank and ate in the party
Google output

rāman pārtti kutikkayuṁ tinnu
Raman party drink-PAST-CONJ eat-PAST
Actual Output
rāman pārttiyil kutikkukayuṁ tinnukayuṁ ceytu
Raman party-LOC eat-PAST-CONJ eat-PAST-CONJ do-PAST

So here the issue also pertains to the case marker in nouns also. When translating to Malayalam, an additional verb form do-PAST is added irrespective any verbs when they occur in conjunction. This should be a handled either in the postposition.

Similarly in interrogative sentences also the interrogation marker is attached to the verb.

(16) Did he come today?
Google output
avan innu vannat?
He today come-PAST-NOML
Actual Output
avan innu vannuvō?
He today come-PAST

In sentence (16) we can see that the verb got the nominalising suffix '*at*' instead of interrogative marker in the sentence translated by google translate.

5 PROPOSED SOLUTIONS FOR THE ISSUES

We know that the SMT system performance depends on the training corpus. The agglutinative language like Malayalam is having many issues pertaining to the corpus. By analysis of verb, we had identified 890 inflections for a single verb based on tense, aspect, modality, interrogation, conjunction, conditionals, person, number and gender. i.e a single verb can generate 890 forms. It may contain noun form, adjectival form and adverbial form. As per Suranad Kunjan Pillai's observation he identified 2880 verbs in Malayalam. Many are rarely used and some are seldom used in current context. By taking in to consideration all these 2880 verbs we can have 2563200 verb forms. We need that much set of parallel corpus to include all these forms. Again this will increase as the language like Malayalam is having compound verbs by combining two verb forms together. A past form of a verb followed by the tense form of particular sentence will form a compound verb as discussed in section 1. Sanskrit nouns combined with the verb form '*āk*' and '*cey*' are also common in Malayalam. But the commonly used verbs will be somewhat less. That may account to around 2000. If we identify all those verbs that are commonly used in Malayalam by the corpus analysis, then we can incorporate this in the glossary and the SMT performance can be improved based on that. Added to these worry we have 8 cases in Malayalam [9]. These eight cases will get doubled and can have 16 forms for noun. As we know that the common noun and proper nouns will be in millions. If we consider all the forms it will be a multitude of 16. So the sparsity becomes much higher in Malayalam for SMT system. This should be handled by some other mechanism instead of adding more parallel corpus by incorporating all the factors.

MOSES MT system introduced a factored MT system in order to handle the data sparsity. But this cannot handle the verb issues that discussed in section 3. The factored translation model take care of lemmas and meanings [7] as shown below

The three mapping steps in our morphological analysis and generation model may provide the following applicable mappings:

- **Translation:** Mapping lemmas
  - *haus -> house, home, building, shell*

- **Translation:** Mapping morphology
  - *NN\plural-nominative-neutral -> NN\plural, NN\singular*

- **Generation:** Generating surface forms
  - *house\NN\plural -> houses*
  - *house\NN\singular -> house*
  - *home\NN\plural -> homes*

A language modelling will take care of the structure of the target language sentences. But if the phrasal mapping is not properly done, then the language model will also fail. We need to have a complete tagged corpus along with their complete syntax and semantic information. Then only that parameter can take in to account for the processing. So we need to analyze the verb first for the proper mapping of the factors. Consider the sentence below:

(17) he had been writing
avan eḻutikkoṇṭirikkunnuṇṭāyirunnu
he write+PROG+PERF+be+ PAST

Here by analyzing the verb, if we get all features then the parameter needed for the target language can be derived and mapped. Malayalam or any similar agglutinative languages needed a perfect morphological analyzer to handle this issue. If we can identify all the possible verb inflections then by using all the verb forms of a single verb can be used for the translation of all other verbs. In Malayalam, Jayan et al.[5] developed a verb classification. These classes can be used for the analysis. Any verb falls in this category can be analyzed and translated using this.
A Factored Language Model (FLM) considers a word as a collection of features or factors, one of which may be the actual surface form of the word. As described by Kirchhoff et al.[10], a word w is a bundle or vector of K (parallel) factors such that

$$w \equiv \{f^1, f^2, \dots f^K\} = f^{1:K}$$

The notation for the factored language model representation of a probability model over a sentence of T words, each with K factors, is:

$$P(w_1, w_2, \dots w_T) = P(f_1^{1:K}, f_2^{1:K}, \dots f_T^{1:K}) = P(f_{1:T}^{1:K})$$

Factors of a word can be anything, including word classes, morphological classes, stems, roots, or any other linguistic feature, such as may be found in highly inflected languages (Bilmes and Kirchhoff 2003). The surface form of a word can be a factor of itself, so the probabilistic language model can be over both words and their decomposition factors. For example, if we have part-of-speech information and we would like to use it as an additional factor, then the factored representation of words would look like:

the = (“the” , article)
black = (“black” , adjective)
cat = (“cat” , noun)
A factored language model, however, does not impose a particular linear ordering on the factors in a word bundle. As such, there is not a fixed sequence of factors upon which to apply the chain rule. One possible factored language model probability for a sentence with T words could be

calculated by taking the word bundles in sentence order, and the word features in the order in which they are arrayed within each bundle:

$$P(f_{1:T}^{1:K}) = P(f_1^{1:K}, f_2^{1:K}, \dots f_T^{1:K})$$
$$= \prod_{t}^{T} P(f_1^{1:K} | f_{t-(n-1)}^{1:K}, \dots, f_{t-1}^{1:K})$$
$$\equiv \prod_{t}^{T} \prod_{k}^{K} P(f_t^k | f_t^{1:K-1}, f_{t-(n-1)}^{1:K} \dots f_{t-1}^{1:K})$$

Furthermore, not all available features have to be used in a factored language model. The relevant history for a word can be defined to be any subset of the available n ·K factors. This might be useful if certain factors in the corpus are known to be unreliable, perhaps because the tools to generate them were unreliable. From these n.K factors that a factored language model can consider, there are 2^{nK} subsets of variables that can be used as the parent factors– the priors for the conditional probability. The factors within each subset can be permuted to form a distinct factor history. As this is the same as sampling without replacement, a set of factors can be used to represent a family of up to distinct factored language models.

$$\sum_{x=1}^{nK} \binom{nk}{x} x! = \sum_{x=1}^{nK} \frac{(nk)!}{(nk-x)!}$$

The factored language model framework thus allows the lexical context to be tailored to the most useful information in the factored corpus. Later in this chapter, we will discuss smoothing methods for estimating previously unseen events. These are another significant way in which factored language models are more powerful than n-gram models.

The underlying language model probability estimate of the likelihood of word wt is calculated according to the model:

$$P(w_t) = P(w_t | s_t, m_t).P(s_t | m_t, w_{t-1}, w_{t-2}).P(m_t, w_{t-1}, w_{t-2})$$

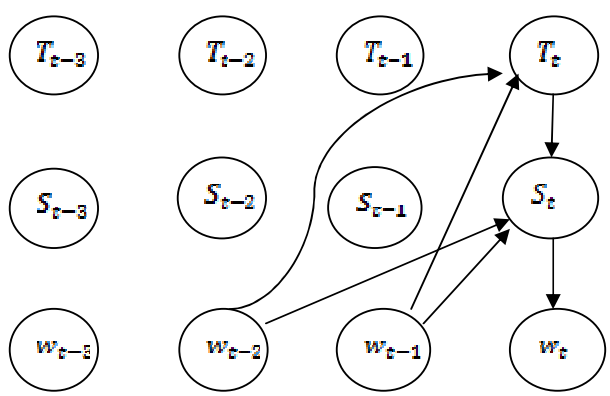

Figure 3.1: Graphical model of a factored language model over words W, stems S, and morphological factors M.

The complete process can be summarized as follows:

a. Analyze the source and target language and find the factors

This will help to find all the features and the equivalents in target language. This will also help to prepare data according to the alignment algorithm.

b. Run word alignment algorithm and align words
This will make use of an alignment algorithm and the word by word mapping table is generated by running the complete parallel corpus.

c. Lexical translation
This stage will estimate the maximum likelyhood lexical translation table in both directions viz, source to target and vice versa.

d. Extract phrases
In this stage the phrases has to be extracted and put in to a single file along with their alignment points.

e. Reordering model
The reordering model mainly depends on the language modelling that is used upon the training set. The phrases that are mapped are reordered based on this language model.

f. Generation model
In the target side the generated text may not be in a proper manner. This can be resolved by using the language rules based on some sandhi processing which are the specialities of the agglutinative language like Malayalam.

6. EXPERIMENTS

Kirchhoff and Yang selected the part of speech (POS) tag and the word stem as additional features for their factored translation model. We followed their example and used the same tools to generate the factored language model features for our factored corpus as well as the development and evaluation sets. Using the tagger and stemmer, we produced a corpus formatted for factored language model training, wherein each English word is expanded to a feature bundle Tag-feature:Tag-feature:... The three features in this corpus are the surface form of the word (W), the part-of-speech tag (T), and the word stem (S), as seen here:

W-he:T-n_nn:S-avan_ W-is:T-v_aux:S-AN W-playing:T-vm:S-kaLikkunnu
This was a naive way of adding factors to the corpus, but a necessary one as the tools used to align and train the corpus, such as GIZA++, do not support factoring words into features.
We used a standard phrase-based statistical machine translation framework for our language model experiments, along with the following software tools:

Pharaoh : The translation models were compiled using Pharaoh, a phrase-based decoder. Pharaoh uses a multi-word phrase translation table along with a language model to translate sentences. Output sentences are produced from left to right, using translation hypotheses selected by a beam search algorithm.

GIZA++ : The translation table was produced by GIZA++, which trains word-based translation models from aligned parallel corpora [12]. GIZA++ is an implementation of the IBM models, so it induces word-level alignments between sentences.

SRILM Toolkit : The n-gram and factored language models were trained using the SRI Language Modeling Toolkit. The improvements to support factored language models were written as part of the 2002 summer workshop at CLSP at Johns Hopkins University (Kirchhoff et al. 2003).

MERT : The translation system tuning was done using the Minimum-Error-Rate Training tool, which is an implementation of the Expectation-Maximization (EM) algorithm. MERT operates by using a pre-calculated language model and set of probabilistic alignments, and then optimizing the weights for the features to maximize the overall system's BLEU score on a reference set.

We used a small training set for the experimental purpose. Lack of sufficient parallel corpus was another major factor. We made use of the corpus available with us which are tagged using the BIS tagset[11]. Although we used a small training set the results were very promising.

We considered 1000 sentence for training and 100 sentences for evaluation. For language modelling we had taken 10000 English sentences. We had analyzed the output based on the BLEU toolkit. Table below shows some results.

Table 3. Evaluation Report

Sl No	Method	BLUE Score
1	Un-factored corpus	12.83
2	POS Tagged corpus	13.01
3	With Morphological Factors	14.24

We got a very low scoring mainly due to the fact that the corpus size is very small. One interesting factor that can be pointed out here is that the performance of the system improves if we incorporate the morphological factors. If we can incorporate large set of trained parallel corpus, then the system will be able to give much better results.

7. CONCLUSIONS

The verb analysis in the agglutinative language like Malayalam of Dravidian language family is in nascent stage. The observations that we have put down here are based on the analysis of google translate and MOSES SMT systems. Basically the verbs in Malayalam carry much information which is helpful in identifying the type of sentence. This will help in the study of language models for finding the prosody pattern to some extent. Normally the SMT is suitable for the translation among the languages of the same family. When it comes to different families the SMT accuracy falls down gradually, especially in the agglutinative languages. Factored translation will account for the semantics of the sentences. The reordering based on the language modelling will help to handle the sentence structure. A further research in identifying all verbs in Malayalam required for completely incorporating the verb forms. This analysis helps to reduce the lexical sparsity for SMT systems. By analysing the English we found that there are not much morphological variations as compared to Malayalam. The Malayalam is having the morphological variations in nouns, pronouns and Verbs in a structured manner. So it is essential to completely analyze all these language units and get the root form for further processing. We tried to focus only on the verb morphology that is the main factor that really affects the quality of the translated sentences.

ABBREVIATIONS/ACRONYMS

PAST – Past Tense, PERF – Perfect Aspect, INFIN – Infinitive, PRES – Present Tense, FUT – Future Tense, PROG – Progressive Aspect, PERMIS – Permission, TRANS – Transitive, CAUS – Causative, ACC – Accusative case, GEN – Genitive case, INSTR – Instrumental case, DAT – Dative Case, NEG – Negative, LOC – Locative, CONJ – Conjuction, MOD – Modal, NOML – Nominalising Suffix

REFERENCES

[1] Malayalam Literary Survey, Volume 27, Kerala Sahitya Akademi, 2005
[2] Suranad Kunjan Pillai, (2000) Malayalam Lexicon, Volume I, Appendix pp 80-81, The University of Kerala.

[3] Mary Priya Sebastian and Dr. G. Santhosh Kumar, Handling OOV Words in Phrase - Based Statistical Machine Translation for Malayalam, CSI Digital Resource Centre, 3rd National Conference on Indian Language Computing organised by Dept. of Computer Applications, CUSAT technically sponsored by Div III, CSI

[4] Mary Priya Sebastian, Sheena Kurian K and Dr. G. Santhosh Kumar, Techniques to Improve the word alignments in Statistical Machine Translation from English to Malayalam, Dyuti, CUSAT digital library(2010)

[5] Ravindra Kumar R, Sulochana K G, Jayan V, Computational Aspect of Verb Classification in Malayalam,I nformation Systems for Indian Languages Communications in Computer and Information Science Volume 139, 2011, pp 15-22

[6] Och, Franz Josef (September 12, 2005), Statistical Machine Translation: Foundations and Recent Advances, The Tenth Machine Translation Summit (PDF), Phuket, Thailand, retrieved December 19, 2010

[7] Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondˇrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, Alexandra Constantin, Christine Corbett Moran, Evan Herbst, Open Source Toolkit for Statistical Machine ranslation: Factored Translation Models and Confusion Network Decoding, Final Report of the 2006 Language Engineering Workshop, Johns Hopkins University Center for Speech and Language Processing

[8] R E Asher and T C Kumari, Malayalam, Descriptive Grammers, Routledge, London and New York, pp – 272-284, 2000

[9] Sunil R, Manohar, N, Jayan, V, Sulochana, K.G, Development of Malayalam Text Generator for translation from English, Annual IEEE India Conference (INDICON), 2011, Page(s): 1 – 6

[10] Bilmes, Jeff A., and Katrin Kirchhoff. Factored language models and generalized parallel backoff. In HLT-NAACL 2003: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada. Association for Computational Linguistics, 2003, Page(s): 4 – 6

[11] Ms. Swaran Lata, Prof. Girish Nath Jha, Dr. Somnath Chandra, Dipti Misra Sharma, Somi Ram, Prof. Uma Maheswara Rao G, Dr. Sobha L, Menak. S, Kalika Bali, Prof. Pushpak Bhattacharyya, Prof. Malhar Kulkarni, Lata Popale, Kirtida Shah, Mona Parakh, Jyoti Pawar, Madhavi Sardesai, Ramnath, Aadil Kak, Nazima, Dr. Richa, Mazhar Mehdi Hussain, Mr. Prashant Verma, Swati Arora, Unified Parts of Speech (POS) Standard in Indian Languages, http://www.tdil-dc.in/tdildcMain/articles/780732Draft%20POS%20Tag%20standard.pdf, Page(s): 17 – 21

[12] Och and Ney, Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, Page(s): 295-302

Authors

Jayan V

Mr. Jayan V is working as Senior Engineer in the Language Technology Centre at C-DAC Thiruvananthapuram, India since October 2005. His areas of interest include Natural Language Processing, Speech Processing, Corpus Linguistics, Information Retrieval and Extraction, etc. He authored and co-authored more than 15 papers in the proceedings of different National and International conferences.

Bhadran V K

Bhadran has been pivotal in establishing the Resource Centre for Cyber Forensics at CDAC Thiruvananthapuram. He has spearheaded the development activities in network forensics and Enterprise Forensics System with advanced capabilities for policy based monitoring and mitigation. He has lead the development work on Stag-analysis, Image forensics and network intrusion analysis.

Currently, he is working as Head, Language Technology leading research in machine translation, automatic speech recognition, text to speech and optical character recognition both printed and handwritten documents. His new area of research interest includes Bionics and Assistive Technology, Natural Interfaces and Autonomous Systems.