IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Natural Language Understanding of Malayalam Language

Usha K[1*], S Lakshmana Pandian[2]

[1, 2] Dept. of Computer Science and Engg, Pondicherry Engineering College, Puducherry-14. India.

Corresponding Author: usha.dileep@pec.edu

DOI: https://doi.org/10.26438/ijcse/v7si8.133138 | Available online at: www.ijcseonline.org

Abstract**—**Natural Language Understanding (NLU) is really challenging sub domain of language processing as far as any highly agglutinative and morphologically rich south Indian Languages are concerned. It requires highly complex procedures and techniques to extract its inflections and grammatical information, thereby make a computer to understand the real sense of the language as human beings does. This paper aims not only providing insights into natural language understanding but also gathers information about the various existing techniques in Malayalam Language. Natural language Understanding refers to the understanding of any language by a machine with the help of an intermediate representation. Here we have briefed the various techniques and algorithms used with Morphological Analyzer, POS tagger, chunking, Parsing, Named Entity Recognition, and Word Sense Disambiguation which are the inevitable components for understanding any Natural Language.

Keywords**—**Natural Language Understanding, Lemmatization, Suffix stripping, Sequence labelling, Stemming, POS tagger, Hidden Markov Model, Support Vector Machine, Morphology, Parsing, Chunking, Sandhi Splitter, Named Entity Recognition, Word Sense Disambiguation.

I. INTRODUCTION

Natural language understanding was defined as the identification of the preconceived semantic from multiple available semantics that could be evolved from any natural language expression which takes the form of well-ordered notation of natural language concept. The grouping of inflected forms of a word as a single item, which can be easily identified by the word's lemma or dictionary form is called lemmatization. Lemmatization depends on correctly identifying the intended part of a text and meaning of a word in a sentence based on its context eliminating ambiguity or even an entire document. The major difference between language processing and language understanding is the notion of finding the real meaning of any sentence based on the context in which it is used.

II. RELATED WORK

Malayalam is a highly agglutinative and morphologically rich language which is rich in inflections. It was spoken by around 37 million people and was derived from Tamil and has been influenced by Sanskrit too. The different approaches used in Malayalam include morphological analysis, POS tagging, disambiguation, chunking and parsing.

2.1 Morphological Analysis

The very basic nature of morphological analysis is the splitting of words into morpheme components. The earlier approaches for implementing a morphological analyser were brute force method, affix driven approach by (PC Kimmo1990) and root driven method (Ample 1998).

2.1.1 Two level morphology

In 1983 Kimmo Koskenniemi has introduced two level morphology for the construction of morphological analyser for Finnish language. In the first level words were described in the way they occur in normal text as surface form and in the second level words were distinguished as stems and suffixes encoded as lexical units.

Two level morphology were based on the given three approaches: Rules were symbol to symbol constraints and have been applied in parallel, not in sequential order. The constraints have to be referred as either surface form or lexical context or as both simultaneously. Lexical look up and morphological analysis has been performed one after another.

2.1.2 Finite State Automata and Finite State Transducer

Finite state automata (FSA) make use of states, transitions and actions. Regular Expressions have been used to accept or reject a string in a particular language. In the initial stage it started accepting character by character thereby shifting from the current state to a new state and when it has reached the final state, stopped its work.

Finite state transducer (FST) is an FSA that worked on two tapes making use of two level morphological concepts. It was a translator which read input from one tape and wrote the translated output on to the second tape. It can be applied on both analyser which makes use of input tape and generator which generates the output onto the second tape along with combining the lexicon and orthographic rules. Its major application is spelling correction of each individual word in a sentence. It is claimed that a Malayalam morphological Analyser using Stuttgart FST is being designed by Santhosh in 2016 which analyses around 150,000 words.

2.1.3 Directed Acyclic Word Graph (DAWG)

DAWG is a very efficient data structure that is used for developing morphological analyser and generator. It is language independent and not using any morphological rules or any other linguistic information. It was also appropriate for lexicon representation and fast string matching. The main reason for using finite-state automata in the NLP domain was that representation of the set of words is highly beneficial and also string look up in a dictionary represented by a FSA is very fast proportional to the length of the string. For the construction of sorted data a data structure named trie has been introduced which is a dictionary with a tree-structured transition graph in which assigns root as the start state and leaves as final states. Dictionary minimization algorithms are totally systematic in terms of the size of their input dictionary. In certain algorithms the memory and time requirements are both linear with the number of states.

In 2001 Kyriakos N. Sgarbas et al., proposed an algorithm which adds a new word to an existing acyclic NFA. In 2003(Kavallieratou) observed that the identification process has been supported by a lexical component based on dynamic acyclic FSAs. Thus acyclic FSAs provides a logical data structure for lexicon representation as well as fast string matching.

2.1.4 Paradigm Approach

Paradigm is the set of all analogous word-forms related to a given lexeme. This approach seemed to be coherent for all inflectionally rich languages. First of all, the possible inflections of a lexical item which can be represented using some paradigm is listed out. Words in the list that belongs to a particular paradigm were classified according to their respective parts-of-speech, noun locatives, post position and preposition. Now paradigm categorization is done based on their morphophonemic manner. Any number of inflection list can be created for each paradigm and there is no limitation on its length too. Similarly any number of paradigms can be assigned to a lexical component, but it will be usually less than five for commonly occurring words. Paradigm approach uses a dictionary which contains each word associated with its paradigm number.

2.1.5 Suffix Stripping

Suffix stripping was another method that makes use of dictionaries and rules to identify a stem and its suffixes. The dictionary to identify a valid stem was called stem dictionary and a second dictionary called suffix dictionary identifies a valid suffix, i.e. it constitutes all suffixes for nouns and verbs in a language. This method also uses morphotactic rules and sandhi rules. Even though the word was not present in the dictionary, the analyser can obtain the stem and its suffixes from the different inflections. Suffix stripping algorithms uses rules to find its root/stem form. Malayalam language is composed of many morphologically complex words which can be obtained by constantly adding affixes to stem. Suffix stripping known to be a simpler approach and its searching time is relatively fast as search was performed on suffixes exclusively. For morphological analysis and morphological parsing of highly agglutinative languages morphosyntactic approaches using finite state mechanism like finite state automata and finite state transducer came into existence. Finite state transducer was used for lexical processing and recursive suffix stripping has been introduced which was useful in the case of compounded words which contain ten or more number of words added in conjunction.

2.1.6 Hybrid Method

Hybrid method has taken the advantage of both Paradigm and suffix stripping approach. Words that belong to noun or verb category are classified into different paradigms. Then apply suffix stripping method to identify valid stem and suffixes. Words are analysed with the inflection list and then use the rule for longest matching suffix present in the suffix list to compare with the available list of suffixes can be used. In 2012 a hybrid methodology has been implemented making use of lttoolbox which in turn make use of FST approach to perform lexical processing. Thus Monolingual, Bilingual and post-generation dictionaries have been used to support noun and verb paradigms of Malayalam with the addition of Lttoolbox.

2.1.7 Probabilistic Method and Rule based Method

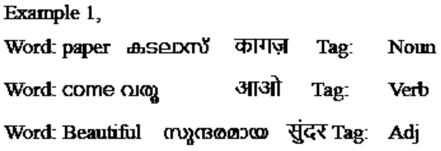
This method presented use of a suffix list and an inflection list. Inflection contains the list of all inflections possible for a particular word and it helped the recognition of the order in which suffixes were attached. The suffix list contains the list of possible suffixes. Morphological analysis has been carried out with two methods: probabilistic and rule based. Probabilistic approach included two methods viz. table look up process and analysis process. In table lookup process, Inflection list has been used for training. During analysis each input word has been segmented into a set of different combinations and valid suffix was identified with help of suffix list. Correct combination was identified from the set with the help of the probability calculated for each suffix.

2.1.8 Corpus based Approach

Corpus based approach has been used for the construction of Morphological Analyser and Generator (MAG) by Antony P J et al., (2012). Here a large sized well generated corpus was used for training a machine learning algorithm and also lot of statistical information has been collected.

2.2 POS Tagging Approaches for Malayalam

The process of assigning any one of the parts-of-speech to a given word is called parts-of-speech tagging (POS Tagging) POS includes nouns, verbs, adjectives, adverbs, prepositions, conjunctions, interjections and their sub category. POS tagger is responsible for POS tagging words in a language.



Example 1,

Word: paper	കടലാസ്	कागज़	Tag:	Noun
Word: come	വരൂ	आओ	Tag:	Verb
Word: Beautiful	സുന്ദരമായ	सुंदर	Tag:	Adj

Mainly POS taggers in Malayalam are classified into rule based tagger and stochastic tagger. Based on contextual information and order of morphemes a large database of hand written disambiguation rules were made in the rule based approach. Stochastic approaches are based on maximum likelihood probability, lexical probability and n-gram probability. Hence Hidden Markov Model followed by Viterbi algorithm has been chosen for most unsupervised models. Manju K et.al has been designed a POS tagger using statistical approach in 2009.

2.2.1 HMM Based Tagging

HMM based tagging can be performed with the help of an annotated corpus. So a tagged corpus was generated from the training set. As Malayalam is a morphologically rich language, the morphological sequences are modeled through deterministic finite automata. When a text has been given as input to the MA, the text is transliterated to an intermediate representation which was used while traversing FSA. Sentences are fed to a tokenizer which compare with dictionary in order to identify the presence of tokens that were valid. Affixes identified based on the morpho-tactics of the root category of the word. A rule based tagger is also used in this architecture to disambiguate the word.

Both unigram and bigram taggers were used here. The automaton has been represented as a directed graph. Also statistical data was collected from the annotated corpus and HMM has been applied to find the maximum probability of every word to assign a particular tag to that word. The perception behind the usage of HMM was to pick the most likely tag for each word. The unigram tagger considers only the probability of the word for any tag given. But for a word sequence HMM tagger has chosen the tag that maximizes the probability with respect to neighbouring tags. For finding maximum probability HMM make use of viterbi algorithm in the case of tagged corpus.

2.2.2 SVM Based Tagging

The tag set proposed for Malayalam language have 29 tags of which 5 tags assigned for nouns, 1 tag for pronoun, verbs were represented by another 7 tags, Punctuation were represented by 3 tags, again adjective, adverb, conjunction, echo, reduplication, intensifier, postpositions, emphasize, determiner, complementizer and question word by 1 tag respectively and finally 2 for numbers. Input was in the form of untagged text and has been manually tagged using the above given tags. Thus obtained manually annotated corpus has been trained using SVM tool and the output thus acquired is a dictionary with merged model and its lexicon. Thus a POS tagger was created which tagged around 1, 80,000 words and achieved an accuracy of 94% as corpus size increases.

2.3 Chunking of Sentences in Malayalam

Chunking is the identification of a group of correlated word from a given raw sentence and reduces the computation. Shallow parser (Hammerton et al., 2002) designed has three modules: tokenizer, POS tagger, chunker in order. Sandhi splitter (Devadath et al., 2014) has been designed that divides a set of conjugated words into a series of individual words in which each word has independent existence. It applied bayesian method at character level to find out the precise split points and has used handy crafted rules to induce morphophonemic changes. The architecture of shallow parser consists of three modules: Sandhi splitter, POS tagger and Chunker. A POS tagger has been built incorporating external and internal word features using CRF (Lafferty et al., 2001) and after doing Sandhi splitting and validation, POS information has also been incorporated. They have used BIS tag set designed by IIIT for Indian languages exclusively. For evaluating the effect of Sandhi in Chunking, a Chunker has been created based on Chunk information using IIIT tag set (Bharati et al., 2006). For shallow parsing sentences a raw input text has been given to the sandhi splitter which identifies individual words and then it was POS tagged and finally fed to a chunker which group the words to form a chunked sentence.

2.4 Parsing of Sentences in Malayalam

Parsing of sentences was important as far the semantic aspect of any language was concerned in NLP application like information retrieval, extraction and question answering. In NLP syntactic parsing means the syntactic analysis of the structure of a text which was made up of a sequence of tokens based on a given formal grammar. A well-known parsing approach named Nivre's Parser has been successfully implemented in many languages like Turkey, English, Hindi and Swedish. Collins and Mc Donalds are other commonly known parsing techniques. Natural language Parsing can be classified into 3 categories: rule based, statistical based and generalized parsers which follow either bottom-up or top-down approach. Bottom-up approach is achieving more popularity when we go for machine learning and deep learning approaches where possible semantics of each word can be considered.

III. NATURAL LANGUAGE UNDERSTANDING OF MALAYALAM - PROPOSED APPROACH

Here we have proposed a Natural Language Understanding approach which takes any language text as input and morphological analysis is performed on each word of input text so as to retrieve the root word. Once the morpheme is obtained it has been analysed and if it is found to be a complex morpheme then it is given to a sandhi splitter which splits the complex word into its component morphemes. Thus we obtain the base form of every word and a part of speech tag was assigned to those words in order to get a tagged corpus. Further, issue of ambiguities are there, then it can be resolved using word sense disambiguation and named entity recognition.

Whenever the real sense of the word is quite visible, POS tagging is performed and then chunking is performed to group the words in a meaningful manner. This is followed by a parsing module which in turn generate a set of meaningful sentences. Parser has been added due to the intention of checking the syntactic structure of a sentence. To make natural language understanding more effective we have added little semantic relatedness and thus a semantic parsing module has been incorporated.

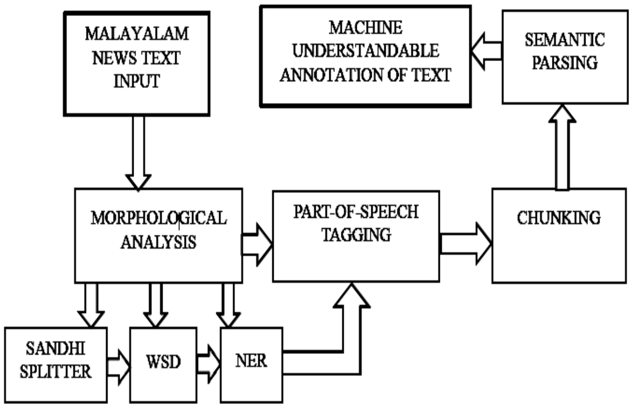
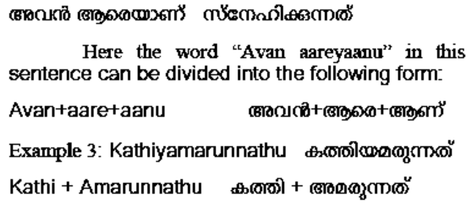


Fig 1: Architecture of Natural Language Understanding for Malayalam Text.

3.1 Sandhi Splitter

In natural languages multiple words may concatenate to yield a single word which may be a complex or a compounded word with certain morphophonemic changes at the point of concatenation. This is known as Sandhi. Sandhi happens in Dravidian languages at two levels. In morpheme level, stem or root join with the affixes to create a word along with morphophonemic changes was considered as Internal Sandhi. In word level, Sandhi between words as in the example 1 given is known as External Sandhi. For example consider the Malayalam sentence equivalent to: Whom he is loving?

Example 2: Avan aareyaanu snehikkunnathu?



അവൻ ആരെയാണ് സ്നേഹിക്കുന്നത്

Here the word "Avan aareyaanu" in this sentence can be divided into the following form:

Avan+aare+aanu അവൻ+ആരെ+ആണ്

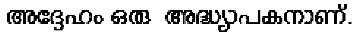
Example 3: Kathiyamarunnathu കത്തിയമരുന്നത്

Kathi + Amarunnathu കത്തി + അമരുന്നത്

After combining the split an extra 'y' gets added to compounded word. This indicates that actual words can be recognized not only by splitting but also by finding the exact split for each compounded word.

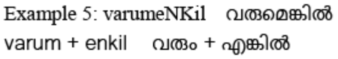
In 2016 Devadath V V et al., created a POS tagger which uses word context level features and evaluated the effect of Sandhi by creating a POS tagger which uses both word-external (context) and word-internal features (morphological features) and the Tag-set used were BIS tag-set2.The POS information has been incorporated only after splitting and validating the Sandhi.

To study the effect of Sandhi on POS, chunks are identified based on the POS tags of words. But if the individual words are not correctly identified, inappropriate POS tags will be assigned and meaningless chunks will be created.
Example 4: addEhaM oru addhyaapakan aaNu.



അദ്ദേഹം ഒരു അദ്ധ്യാപകനാണ്.

There are 4 words and 3 chunks in the above example 3, [addEhaM PRP] NP, [oru QT addhyaapakan NN] NP, and [aaNu VM] VP. If the Sandhi is not identified and individual words are not extracted, system will fail to identify the meaningful sub-parts of a sentence like constituents.



Example 5: varumeNKil വരുമെങ്കിൽ

varum + enkil വരും + എങ്കിൽ

Similarly in example 4, the string varumenkil has two words and two chunks [varuM VM] VP and [enkil CC] CCP. Hence processing Sandhi in the first stage is extremely important in any NLP task for Dravidian languages.

3.2 Word Sense Disambiguation

Word sense Disambiguation (WSD) is a known linguistic mechanism for automatically defining the exact sense of an ambiguous word based on its context. Identifying the actual meaning of a word in a particular context from a collection of meanings of same word at different contexts was assumed as a long standing problem in computational linguistics. In other words WSD is the problem of selecting a sense for a particular word from a set of predefined possibilities. Applications of natural language understanding where WSD is considered includes Knowledge representation, Text mining, Information retrieval, Information extraction and Machine Translation.

WSD can be represented through knowledge base approach or machine learning approaches which can be supervised and unsupervised. Knowledge base approach was easy to implement because they don't use any corpus based training and hence only require a look up on external lexical resources viz. dictionaries, WordNet. But machine learning algorithms need corpus and training. In 2010 Haron et al., make use of a knowledge base approach for Malayalam language. Two algorithms have been designed by them and former was based on Lesk and Walkers algorithm whereas the later was based on Conceptual density which takes sematic relatedness of words into consideration. In Coling (Jayan V, 2012) et al., proposed disambiguation of prepositions in English language. It is seen that prepositions in English is replaced by post positions in almost all Indian languages. Post positions in Malayalam are free form which comes immediately after nouns and establish a sort of grammatical relationships between nouns and verbs of sentences. Sheshagiri prabhu decribes postpositions as words added to case suffixes which signify special meaning and had listed around 80 postpositions in Malayalam.

Example 6: Njaan innale oru puthiya kathi vaangi.

ഞാൻ ഇന്നലെ ഒരു പുതിയ കത്തി വാങ്ങി

Ellam theeyil kathiyamarunnathu avar nokkininnu.

എല്ലാം തീയിൽ കത്തിയമരുന്നത് അവർ നിസ്സംഗതയോടെ നോക്കി നിന്നു

The example given above illustrates the need of sandhi as well as WSD for Malayalam language. Here the word 'kathi is used in two senses and also it make use of two POS tags. The meaning of 'kathi' in first sentence is 'knife' where as in second one it means 'burning'. So ambiguity resolution, sandhi splitting (കത്തിയമരുന്നത്) and NER is needed in theses sort of sentences.

3.3 Named Entity Recognition

Named Entity recognition (NER) is also known as entity chunking, entity identification and entity extraction. Named entities are nouns which are difficult to identify and even after identification it is tedious to classify. NER is commonly implemented using three approaches viz. rule based, machine learning based and a hybrid of both. The machine learning models comprises supervised and unsupervised learning approaches for NER using statistical methods like HMM, Decision Forest, Maximum Entropy, SVM and CRF.

In 2013 Jisha P Jayan et al., proposed a hybrid statistical approach for NER in Malayalam. Here the Named Entity hierarchy is divided into Entity Name, Time and numerical expression and then tagged using Indian language Machine language Tagset. For example KAVITHA denotes a name of person as well as poem in Malayalam literature. The results shows that a hybrid approach is more suited than using machine learning or rule based approach individually. In 2014 Pragisha K et al., implemented an NER system as a part of her Question Answering project in Malayalam. The NER system was based on a hybrid approach on dictionary based extraction and regular expression method. In 2015 Sanjay S P et al., CRF based NER was for twitter micro post were implemented for English, Malayalam Tamil and Hindi. The features extracted from words present in the tweets are trained using CRF which generates unigram and bigram features. Here training and testing data are POS tagged with tagger tools by taking POS tagger (Gimbel for English) from respective languages. POS taggers are used to improve the accuracy of named entities recognized. Clusters are taken for English (brown cluster) alone whereas the non-availability of cluster tool was a draw back for Indian languages. Binary features are extracted for those languages and combined as a BIO file which is then tested. Approximate match metric are used for evaluating partial correctness of the named entity.

3.4 Semantic Parsing

The notion of understanding is covered in the parsing of any language. As far as NLU approaches can be classified into four different category. They are:

Distributional approach deals with the application of machine learning and Deep Learning which in turn convert the content into word vectors and thus perform POS, dependency parsing and semantic relatedness.

Frame based approach uses a data structure for representing knowledge. Model-theoretic approach uses linguistic concepts like model theory and compositionality. Interactive learning come into picture where humans taught computers gradually with the help of interactive environment. In this context we are working on a distributional approach in order to perform dependency parsing.

IV. CONCLUSION AND FUTURE WORK

This paper gives a clear idea about the existing natural language processing approaches in Malayalam. The sufficient and necessary modules in order to perform NLP are explained. The idea NLU has been included which necessitates the inclusion of Sandhi splitter, NER, WSD and Semantic Parsing modules. We hope that the proposed NLU approach will be highly beneficial not only for Malayalam but also for other south Indian languages and will open new doors of improvement in language Processing.

Extraction and understanding of information is very difficult from a natural language and hence the problem is classified as NP hard. The application of DEEP learning in the Natural language processing techniques using RNN (Recurrent Neural Network) and LSTM (Long-Short Term Memory) is found to be more efficient and gives a better accuracy that any other machine learning approaches mentioned so far.

REFERENCES

[1] Kavallieratou, Ergina, et al. "Handwritten word recognition based on structural characteristics and lexical support." *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* IEEE, 2003.

[2] Hammerton, James, et al. "Introduction to special issue on machine learning approaches to shallow parsing." *Journal of Machine Learning Research* 2.Mar (2002): 551-558.

[3] Haroon, Rosna P. "Malayalam word sense disambiguation." *2010 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2010.

[4] Aasha, V. C., and Amal Ganesh. "Rule Based Machine Translation: English to Malayalam: A Survey." *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*. Springer, New Delhi, 2016.

[5] Antony, J. Betina, and G. S. Mahalakshmi. "Named entity recognition for Tamil biomedical documents." *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]*. IEEE, 2014.

[6] Nisha, M., and PC Reghu Raj. "Malayalam morphological analysis using MBLP approach." *2015 International Conference on Soft-Computing and Networks Security (ICSNS)*. IEEE, 2015.

[7] Dinh, Phu-Hung, Ngoc-Khuong Nguyen, and Anh-Cuong Le. "Combining statistical machine learning with transformation rule learning for Vietnamese word sense disambiguation." *2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future*. IEEE, 2012.

[8] Khalil El Hindi, Muna Khayyat & Areej Abu Kar (2017) Comparing the Machine Ability to Recognize Hand-Written Hindu and Arabic Digits, Intelligent Automation & Soft Computing, 23:2, 295-301, DOI: 10.1080/10798587.2016.1210257.

[9] Sharma, Sanjeev Kumar, and G. S. Lehal. "Improving Existing Punjabi Grammar Checker." *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*. IEEE, 2016.

[10] Manju, K., S. Soumya, and Sumam Mary Idicula. "Development of a POS tagger for Malayalam-an experience." 2009 International Conference on Advances in Recent Technologies in Communication and Computing. IEEE, 2009.

[11] Idicula, Sumam Mary, and Peter S. David. "A morphological processor for malayalam language." *South Asia Research*27.2 (2007): 173-186.

[12] Rajeev, R. R., and Elizabeth Sherly. "Morph analyser for malayalam language: A suffix stripping approach." Proceedings of 20th Kerala Science Congress, 2007.

[13] Jayan, Jisha P., R. R. Rajeev, and S. Rajendran. "Morphological analyser for malayalam-a comparison of different approaches." *IJCSIT* 2.2 (2009): 155-160.

[14] Vinod, P. M, V. Jayan, and V. K. Bhadran. "Implementation of Malayalam morphological analyzer based on hybrid approach." *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012)*. 2012.

[15] Rinju O.R, Rajeev R R, Raghu Raj P C, Elizabeth Sherly,"Morphological Analyzer for Malayalam: Probabilistic Method vs Rule Based Method", *International Journal of Computational Linguistics and Natural Language Processing, Vol 2 Issue 10 October 2013.*

[16] Jancy Joseph et.al. "Rule based Morphological Analyser for Malayalam nouns", Computational Analysis of Malayalam Linguistics, *IJIRCCE* vol3, special issue 7, OCT 2015.

[17] Nair, Latha R. "Language Parsing and Syntax of Malayalam Language." *2nd International Symposium on Computer, Communication, Control and Automation*. Atlantis Press, 2013.

[18] Jayan, Jisha P., and R. R. Rajeev. "Parts Of Speech Tagger and Chunker for Malayalam–Statistical Approach." *Computer Engineering and Intelligent Systems* 2.2 (2011): 68-78.

[19] Manju, K., S. Soumya, and Sumam Mary Idicula. "Development of a POS tagger for Malayalam-an experience." *2009 International Conference on Advances in Recent Technologies in Communication and Computing*. IEEE, 2009.Latha R Nair, David s Peter "Shallow Parser for Malayalam Language Using Finite State Cascades", *4th International Conference on Image and Signal Processing* 2011.

[20] Nair, Latha R., and S. David Peter. "Development of a rule based learning system for splitting compound words in malayalam language." *2011 IEEE Recent Advances in Intelligent Computational Systems*. IEEE, 2011.

[21] Antony, P. J., Santhanu P. Mohan, and K. P. Soman. "SVM based part of speech tagger for Malayalam." *2010 International Conference on Recent Trends in Information, Telecommunication and Computing*. IEEE, 2010.

[22] Pragisha, K., and Dr PC Reghuraj. "A Natural Language Question Answering System in Malayalam Using Domain Dependent Document Collection as Repository." *International Journal of Computational Linguistics and Natural Language Processing* 3.3 (2014): 2279-0756.

[23] Jayan, Jisha P., R. R. Rajeev, and Elizabeth Sherly. "A hybrid statistical approach for named entity recognition for malayalam language." *Proceedings of the 11th Workshop on Asian Language Resources*. 2013.

[24] Devadath, V. V., and Dipti Misra Sharma. "Significance of an accurate sandhi-splitter in shallow parsing of dravidian languages." *Proceedings of the ACL 2016 Student Research Workshop*. 2016.

AUTHORS PROFILE

I am Usha.K currently pursuing research under Quality Improvement Programme(QIP), at Dept. of Computer Science and Engineering, Pondicherry Engineering College, Puducherry. I did Master of Engineering in Computer Science at PSG college of Technology, Coimbatore. And did my Bachelor of Engineering in Computer Science at Indian Engineering College, Vadakkangulam. I am on deputation from NSS college of Engineering, Palakkad, working as Associate Professor for a period of 15 years. My area of interest includes Natural Language Processing, Machine Learning, Deep Learning and Pattern Recognition.