*Research Journal of Recent Sciences* ISSN 2277-2502
Vol. **3(3),** 99-103, March **(2014)** *Res.J.Recent Sci.*

*Review Paper*

# A Short Review about Manipuri Language Processing

**Surjit Singh R.K.[1], Gunasekaran S.[1], Anand Kumar M.[2] and Soman K.P.[2]**
[1]CSE Department, Coimbatore Institute of Engg and Technology Coimbatore, INDIA
[2]Centre for Excellence in Computational Engg and Networking, Amrita Vishwa Vidyapeetham, Coimbatore, INDIA

**Available online at: www.isca.in, www.isca.me**
Received 13th July 2013, revised 13th August 2013, accepted 26th September 2013

## Abstract

*Manipuri is a highly agglutinating and compounding language. Words in Manipuri language are formed by affixation. New words are formed by appending prefix and suffix to the root word. So, Manipuri Language processing helps in identifying various class of a word in a sentence. Besides this various application and analysis for Manipuri language such as Part of speech Tagging, Morphological Analyzer, Name Entity Recognition, Multiple Word Expressing etc. can be performed easily which is required for Machine Translation. This study presents the review about some of the existing Manipuri language processing tools and their approaches.*

**Keywords:** Manipuri, language processing, morphology, machine translation, part-of-speech tagger.

## Introduction

Natural Language Processing (NLP) is an emerging domain in present scenario for research which leads to the development of our regional language. The motive behind NLP is to educate people which are unable to access the latest technology being developed. NLP includes various computational and analyzing processes which enable machine to understand the language. Every language in the world has its own family. Thus Manipuri language belongs to Tibeto-Burman language. Manipuri language is used as a means of communication, in the neighboring states as well as neighboring countries like Mayanmar and Bangladesh. Among all Tibeto-Burman languages, it is the first language which includes in the Indian Constitution. Manipuri language has its own script and literature and it uses two scripts while writing i.e. Meitei Mayek which is its own original script and Bengali script which is borrowed from Bengali language. It is very difficult to classify or identify word class in Manipuri language as it is highly agglutinative, monosyllabic. Some of the applications for NLP are Part of Speech tagging (POS), Name Entity Recognition (NER), and Multiple Word Expression (MWE) etc. which are used in Machine Translation.

Manipuri language is a less computerized and there is not much work available in web as compared to the language like English, Chinese, and Korean etc. Every now and then there is a great influence of Korean movies in Manipur. This will leads to degradation of Manipuri language. In Manipuri language, designing a model for segmenting a word into syllabic unit is very important for a less computerized language. Segmenting a word into syllabic units helps in improving the application for NLP such as morphology analyzer, text to speech conversion, speech conversion works, lexicon development, spell checker etc. This can be prevented by developing an efficient Electronic-Dictionary which can help learning and teaching process of Manipuri language smoothly. So research in this language processing will help to bring the language globally and it will also help other non Manipuri speaking people to understand the language and brought it up to the global platform. For analyzing purpose, it is not necessary to stored separate text corpora; in fact it should be the text corpora which are used in our day to day life. Manipuri text corpora are being collected from the various news paper and books written in Manipuri language. In this paper, the subsequent section presents the work which has been done by many researchers in Manipuri language.

## Challenges in Manipuri language processing

Manipuri is an agglutinative, monosyllabic, and compounding language. It means that in Manipuri language new words are formed easily and monosyllabic means that even a single letter forms a word which is a meaningful. So considering all these, there are some challenges occurred during the processing of Manipuri Language (ML). Some of these challenges are shown below[1]: i. There is lack of Linguistically Studied (LS) in Manipuri language. That means the word categories in Manipuri language are not well defined. Grammatically, in some cases there is a difference between structural and contextual meaning of a particular word. ii. As compared to the language like, English, Hindi, Chinese, Korean etc., the Resource of the Manipuri language is very less. Here resource represents the text which is machine readable. iii. In order to make a language machine readable, there should be an enough tools, operating system or an application such as processor, translator, compiler, and encoding, decoding support in order to make the particular language computerized. iv. For any language, the processing of language in machine is very complex. So in order to make a machine (computer) understand the language, Language Processing should be properly analyse and effectively computerized.

Apart from all these challenges mentioned above, there is still some area lacking behind which is required for processing Manipuri language. They are Finance, Equipment, the People resource, the Society, the Government, and the political support.

## Language processing tools for Manipuri

There is not much work done in Manipuri language processing but some of the work has been found in the record. This section addresses the tools and technique used in Manipuri Language Processing.

**Part-of-Speech Taggers for Manipuri Language:** Part of speech tagging is an essential phase in natural language processing. It is the process of assigning a tag for an individual word in a sentence corresponding to the part of speech based on definition as well as its context. This is one of the important stages in the field of natural language processing (NLP) which makes machine able to identify the words and its neighboring words in a sentence. The part of speech tagging is used in various applications like information extraction, shallow parsing, and machine translation etc. POS tagging in Manipuri language are performed using rule based as well as statistical based approaches.

Kh Raju Singha et.al presented Manipuri language rule based POS tagging[2], where hand written linguistic rules for Manipuri language are used by applying a technique called affix stripping. It means extracting of prefix and suffix from root words. Based on ILPOST framework a three tier tagset for Manipuri language is designed. A total number of 97 tags including generic attributes and language specific attribute values are used for testing. It applied 25 rules in this system and gives an accuracy of 85% is obtained for 1000 words. In this type of POS tagging accuracy level increases with increase of number of words in lexicon and number of rules applied. POS tagging serves as an interface for morphological analyzer and chunking[3]. So, the output of this system can be used as a corpus in various computational processes of POS tagging for Manipuri Language.

Thoudam Doren Singh and Sivaji Bandyopdhyay[4] developed Manipuri based Morphology POS tagger using handcrafted rules, in which the contextual meaning of the word is not used, rather it uses three dictionary i.e., root dictionary, prefix dictionary as well as suffix dictionary as a feature for morphological POS tagger. A total number of 3784 sentences containing 10917 words are being tested using 13 tagset. This gives an accuracy of 69% in which 31% were incorrectly tagged, 23% were unknown words and 8% of known words are tagged wrongly. The result of morphology POS tagging gives an asset in other approaches of POS tagging in Manipuri Language Processing.

Thoudam Doren Singh et.al developed Manipuri based POS tagger[5] using Conditional Random Field (CRF) which is one of the statistical learning model used in Natural Language Processing (NLP). CRF is defined as a process of statistical modeling method applied in pattern recognition and machine learning. Unlike rule based tagging, CRF based POS tagger used the features of words like contextual text and orthographic word level. Using this method a total number of 63,200 tokens have been manually annotated using 26 tagset which is defined for the Indian language. In this approach the CRF based system is train and tested with the token number of 39449 and 8672 word forms respectively by considering contextual and orthographic word level as feature. After evaluating an accuracy of 72.04% is obtained.

Thoudam Doren Singh et.al has also developed Manipuri based POS tagger[5] using Support Vector Machine (SVM) which is a popular supervised machine learning approach for classification, regression, and other learning task. SVM is introduced by Vapnik. The advantages of SVM is that it is robust, gives high accuracy with large data sizes without over fitting and also helps for easy text categorization. SVM based POS tagger consists of two phase i.e. training and classification phase. YamCha toolkit is used for training an annotated data and TinySVM-0.07 is used for classification. In this technique a total number of 63,200 tokens have been manually annotated using 26 tagset which is defined for the Indian language. Considering different contextual and orthographic word level as a feature, the SVM based system is trained and tested with a token number of 39449 and 8672 word forms respectively. After evaluation, the result obtained an accuracy of 74.38%. Thus SVM based tagger outperforms as compared to the CRF based tagger by a margin of 2.34%.

Kishorjit Nongmeikapam and Sivaji Bandyopadhyay developed Manipuri based POS tagger[6] using Support Vector Machine (SVM) by identifying reduplicated multi word expression (RMWE) as a feature for Manipuri POS tagger. Here in this approach experiments are performed in two phases. In the first phase RMWE are identified using SVM system as well as common feature like surrounding words, stem, number of acceptable suffixes, prefixes, acceptable suffixes, prefixes, length of the word, word frequency, digit and symbol features. In the second phase POS tagging starts using the identified RMWE as well as dynamic POS (i.e. POS of the previous words are considered) as a feature. A total number of 25,000 words are divided into two files i.e. training files and testing files. The testing file consists of 20,000 words and testing files consists of 500 words. The evaluation result obtained an accuracy of 71.15% Recall, 83.15% Precision and 76.68% F-measure which is reasonably significant.

Kh Raju Singha et.al developed Manipuri POS tagger[7] using Hidden Markov Model (HMM) which is a stochastic model and used to solve classification problem that have an inherent state sequence representation. It also uses a little amount of knowledge about the language apart from the simple contextual information. In HMM technique for Manipuri POS tagging, the manually annotated test set data from 97 morpho-syntactic tagset[5] of Manipuri language including generic attribute and tagged corpus have been used. It gives an accuracy of 92% for 2000 tagged lexical item and accuracy increases with the increase of number of tagged corpus. As compared to the result of manually tagging[7], 80% result was found to be correct for the automatically generated sequence test set.

**Morphology Analyzer for Manipuri Language:** Morphology is defined as the process of studying how words are composed of smallest meaning bearing units of the language. This smallest meaning bearing of a language is called as Morphemes. Morphemes are divided into two classes. First one is stem, is the main morpheme of a word which gives meaning to a word. Second one is affixes which are words that combine with stem to modify their meanings and grammatical functions. Manipuri language being agglutinative, there are number of affixes which can easily formed new words. So it is necessary to analyze the words morphologically in order to make it machine readable.

Thoudam Doren Singh and Sivaji Bandyopadhyay presented Manipuri[8] based morphology analyzer using a Manipuri-English-Dictionary. In this approach word class are identified by using affixes (prefix and suffix) attached with a word like noun, verb, adjective, and adverb. Not only this sentence type like, imperative, interrogative and negation etc. can also be identified by suffixes attached to the verb word. Word class and sentence type identification are evaluated using a Morphological analyzer and result obtained is reasonable compared to the human facts.

Thoudam Doren Singh and Sivaji Bandyopadhyay developed Manipuri based Morphology analyzer[9] using a Manipuri-English dictionary which can identify morphemes form raw text. Manipuri sentences are given as an input to the system where for each words produces the root word, the suffixes and the prefixes and English equivalent pattern for the surface level word. This analyzer can also analyze five different types of words, they are: word without any affix, word with a prefix, word with one or more suffix, compound words, reduplicative words. The main purpose of the morphology analyzer is to provide strong platform for machine translation which is a core technique for NLP. So far the developed analyzer is limited and cannot handled colloquial features, phrases and idioms and in situation like repetition of two symbolic characters. The accuracy in this type of analyzer can be improved by adding some specific rules such as feature extraction.

**Name Entity Recognition for Manipuri Language:** From the name itself one can understand the meaning of Name Entity (NE). It helps to recognize the categories of words such as name of the persons and organizations, location, time, date and currency. The advantages of the NER are that it helps machine to differentiate between different object since every object has its own unique names. Some of the problems faced during the process of NER are: less capitalization, lexical are long and word forms are complex, difficult to identified subject and object in a sentence. The different tools and technique are found for NE in Manipuri language.

Kishorjit Nongmeikakpam et.al developed Manipuri Name Entity Recognition[10] using CRF approach, where the process of feature selection was done through manual assumption. After selecting the best feature and experimenting, the result obtained is 81.12% of Recall, 85.67% of Precision, and 83.33% of F-Score.

Thoudam Doren Singh et.al developed Manipuri language NER[11] using SVM based technique. In this technique, Manipuri news corpus is manually annotated using different contextual information of the words and orthographic feature. A total number of 174,921 untagged word-forms corpuses have been manually annotated using coarse-grained tagset containing four Named Entity tags and a best feature of NE have been selected from the untagged word form. Then a token of 28,629 and 4762 word forms have been trained and tested which demonstrated an accuracy of 93.91% of Recall, 95.32% of precision, 94.59% of F-Score.

**Multi Word Expression for Manipuri Language:** Multi Word Expression (MWE) can be defined as minimal unit word in the lexicon of a language, example ‘go’ and ‘went’ and ‘gone’ are all members of the English word ‘go’. In MWE words are composed independently and carry different meaning. For a large scale processing for a language linguistically and machine readable, the technique called multi word expression was developed. Some tools and technique used for identifying MWE in Manipuri language have been developed and were found in record.

Kishorjit Nongmeikakpam et.al presented Conditional Random Field (CRF) based MWE identification technique[12] for Manipuri language. Using MWE technique an accuracy of 60.39% Recall, 85.53% Precision, 70.83%, F-Score was obtained. In Manipuri language, many new words are formed by appending affixes. So by identifying reduplicated words in CRF technique, the accuracy for identifying MWE was found to be further improved from the previous result i.e. 62.24% Recall, 86.06% Precision, 72.24% F-Score.

Thoudam Doren Singh and Sivaji Bandyopadhay developed Manipuri[13] based SVM approaches for identifying MWE. SVM technique is performed by collecting four and half million Manipuri corpora from a popular Manipuri News agency. In this approach, identifying reduplicated words is used as feature and the result is improved significantly. From this corpora using rule based approach, a total number of 28,629 word-forms is manually annotated and 4,763 word-forms are trained and tested, which gives an accuracy of 94.24% Recall, 82.27% Precision, 87.68% F-Score. While with the same data size applying SVM technique an accuracy of 94.62% Recall, 93.53% Precision, and 94.07% F-Score is obtained. So from this result we can clearly see that SVM approaches outperformed rule based technique.

**Machine Translation System for Manipuri Language:** The process of translating the source language text into target language text is called Machine Translation. For a good Machine Translation (MT) system there is some information required about language such as words, their meaning, concept, relative words in another language. So here the main resource is a machine readable electronic dictionary. MT enables different people to understand different language easily. Some of the challenges are there while processing the statistical machine translation from English-Manipuri language like wide syntactic divergence and richer morphology and case marking of Manipuri compared to English. Manipuri language being a less privileged, less computerized; there is not much work available in record for Machine Translation (MT).

Thoudam Doren Singh and Sivaji Bandyopadhyay developed Manipuri based machine translation[14] using morpho-syntactic and semantic information. Here the morphology and dependency relation plays an important role for improving accuracy. The main motive in this approach is to find out fluency and adequacy. Due to restricted bilingual translation from English to Manipuri language, the process becomes a difficult task. In this method the important translation factors considered is the role of suffixes and dependency on the source side i.e. English language and case markers on the target side i.e. Manipuri language. For training purpose a total number of 10350 sentences have been collected from news domain and 500 sentences have been tested for system. After evaluating in this approach, it is found that shorter sentences obtain greater accuracy than the larger sentences in terms of fluency and adequacy.

Thoudam Doren Singh and Sivaji Bandyopadhyay developed Manipuri based SMT[15] using morphology and dependency relation. In this approach, Manipuri language is in source side while English language is in target side in the translation process. Unlike English-Manipuri MT, the important factored consider in this approach is that the role of case markers and POS tags information are at the source side and suffixes and dependency relations are at the target side, but morphological information and semantic relations are incorporated in order to improve output. For finding out fluency and adequacy, subjective evaluation is being conducted. Automatic scoring technique BLEU and NIST are conducted for evaluating purpose where an accuracy of 13.425 baseline BLEU score and 17.537 factored BLEU score are obtained which is a statistically significant improved. The evaluation result shows that shorter sentences are better than longer sentences using semantic relations.

Thoudam Doren Singh and Sivaji Bandyopadhyay have developed the Manipuri-English machine translation[16] based using Manipuri-English example based machine translation system. From news corpora sentence level parallel corpus is built where phrase alignment is performed by applying POS tagging, morphology analysis, NER and chunking. In a situation like word level mismatch, the unmatched target phrase translation are identified and then recombined with the retrieved output. In this approach, EBMT system method is evaluated where an accuracy of 0.137 BLEU and 3.361 NIST is obtained which improved significantly than the baseline SMT system with same training and test data.

**Discussion:** Apart from these language processing tools, annotated corpora are also an essential data for Linguistic research. In automatic language teaching tools, annotated corpus plays an important role[17]. The table 1 shows the different Manipuri language processing tools developed by various authors using different approaches.

**Table-1**
**Existing Manipuri language processing tools**

| Author | Tools | Method |
|---|---|---|
| Thoudam Doren Singh et.al [2] | Part-of-Speech tagger | Hand crafted rules |
| Kh Raju Singha et.al [4] | Part-of-Speech tagger | Rule Based Approach |
| Thoudam Doren Singh et.al [5] | Part-of-Speech tagger | CRF and SVM |
| Kh Raju Singh et.al [7] | Part-of-Speech tagger | Hidden Markov Models |
| Kishorjit Nongmeikakpam et.al [10] | Named Entity Recognition | Conditional Random Fields |
| Thoudam Doren Singh et.al[11] | Named Entity Recognition | Support Vector Machines |
| Kishorjit Nongmeikakpam et.al [12] | Multi-Word-Expressions | Conditional Random Fields |
| Thoudam Doren Singh et.al[13] | Multi-Word-Expressions | Support Vector Machines |
| Thoudam Doren Singh et.al [14] | Machine Translation | Morpho-Syntatic and Semantic Information |
| Thoudam Doren Singh et.al[15] | Machine Translation | Morphology and Dependency relation |
| Thoudam Doren Singh e.al [16] | Machine Translation | Manipuri-English Example Based |

## Conclusion

In this paper some of the existing Manipuri Language Processing tools and their developing methodologies are surveyed. Manipuri language being less computerized there are still resources, tools and techniques to be improved such as increasing the annotated corpora, dictionary, inflection list spelling checking technique for implementing lexical rules, ambiguity and disambiguation scheme since new words can be formed easily by affixing technique. In future it is needed to develop the technique for identifying NER and MWE for improving POS tagging which is useful in MT and hybridization of rule base approach and statistical approaches. So, considering all the above mentioned challenges, it is necessary to developed cross lingual information retrieval system and machine translation for ensuring Manipuri language a highly valued language in the near future.

## References

**1.** Anil Kumar Singh, Language Technologies Research Centre, IIIT, Hydrabad India, NLP for Less Privileged Languages: Where do we come from? Where are we going? *In IJCNLP Workshop on NLP,* **(2008)**

**2.** Kh Raju Singha, Bipul Syam Purkayastha, and Kh Dhiren Singha, Part of Speech Tagging in Manipuri: A Rule-based Approach, *IJCA,* **51(14)**, **(2012)**

**3.** Dhanalakshmi V., Anandkumar M., Shivapratap G., Soman K.P. and Rajendran S., Tamil POS tagging using linear programming, *International Journal of Recent Trends in Engineering*, **1(2)**, 166-169 **(2009)**

**4.** Thoudam Doren Singh, Sivaji Bandyopadhyay, Morphology Driven Manipuri POS Tagger, *In proceeding of IJCNLP-08 Workshop on NLP Hydrabad, India,* **(2008)**

**5.** Thoudam Doren Singh, Asif Ekbal, Sivaji Bandyopadhyay, Manipuri POS Tagging Using CRF and SVM: A Language Independent Approach, In Proceeding of ICON 2008: *6th International Conference on Natural Language Processing* **(2008)**

**6.** Kishorjit Nongmeikapam, Sivaji Bandyopadhyay, "SVM Based Manipuri POS Tagging Using SVM Based Identified Reduplicated MWE (RMWE), *In Proceeding of the CUBE International Information Conference, CUBE,* **(2012)**

**7.** Kh Raju Singha, Bipul Syam Purkayasha, and Kh Dhiren Singha; Part of Speech Tagging in Manipuri with Hidden Markov Model; *International Journal of Computer Science Issues,* **9(6), No 2**, **(2012)**

**8.** Thoudam Doren Singh, Sivaji Bandyopadhyay, Word Class and Sentence Identification in Manipuri Morphological Analyzer, *In Proceedings of MSPIL, IIT Bombay,* **(2006)**

**9.** Thoudam Doren Singh, Sivaji Bandyopadhyay, Manipuri Morphological Analyzer, *In Platinum Jubilee International Conference of the LSI, Hydrabad,* December **(2008)**

**10.** Kishorjit Nongmeikakpam, Leisram Newton Singh, Tontang Shangkhunem, Bishworjit Salam, Chanu, Sivaji Bandyopadhyay, CRF Based Name Entity Recognition in Manipuri: A Highly Agglutinative Indian Language. *In Proceedings of 8th International Conference on Natural Language, IIT Kharagpur,* India, **(2011)**

**11.** Thoudam Doren Singh, Kishorjit Nongmeikakpam, Asif Ekbal, Sivaji Bandyopadhyay; Name Entity Recognition for Manipuri Using SVM, *In Proceedings of Pacific Asia Conference on Language, Information and Computation, Hong Kong*, **(2009)**

**12.** Kishorjit Nongmeikakpam, Sivaji Bandyopadhyay, "Identification of MWE using CRF in Manipuri and Improvement using Reduplicated MWE, *In Proceedings of ICON-2010, IIT Kharagpur, India,* **(2010)**

**13.** Thoudam Doren Singh, Sivaji Bandyopadhay, Web Based Manipuri Corpus for Multiple NER and Reduplicated MWE Identification Using SVM, In *23rd International International Conference on Computational Linguistic (COLING),* Beijing, August **(2010)**

**14.** Thoudam Doren Singh, Sivaji Bandyopadhay, SMT of English-Manipuri using Morpho-syntactic and Semantic Information, *In Proceeding of 9th Conference of the Association for Machine Translation in America (AMTA, 2010),* Denver, Colorado, USA, **(2010)**

**15.** Thoudam Doren Singh, Sivaji Bandyopadhay, Manipuri-English Bidirectional SMT systems using Morphology and Dependency Relations, *In Proceeding of Syntax and Structure in Statistical Translation (SSST-4) of 23rd International Conference on Computational Linguistics (COLING), Beijing,* August **(2010)**

**16.** Thoudam Doren Singh, Sivaji Bandyopadhay, Manipuri-English Example Based Machine Translation System, *IJCLA (ed.) ISSN 0976-0962*, **(2010)**

**17.** Dhanalakshmi V. and S. Rajendran, Natural Language processing Tools for Tamil grammar Learning and Teaching, *International journal of Computer Applications (0975-8887)* **8(14), (2010)**