Transliteration of CRF Based Multiword Expression (MWE) in Manipuri:

From Bengali Script Manipuri to Meitei Mayek(Script) Manipuri

Kishorjit Nongmeikapam[1] , Ningombam Herojit Singh[1], Bishworjit Salam[1],Sivaji Bandyopadhyay[2]
[1] *Dept. of Computer Science and Engg., Manipur Institute of Technology, Manipur University, Imphal, India*
[2] *Dept. of Computer Science and Engg., Jadavpur University, Jadavpur, Kolkata, India*

Abstract—**This paper deals about the transliteration of the identified Multiword Expression (MWE) of Manipuri using Conditional Random Field (CRF). Manipuri is a very highly agglutinative language and is an Eight Scheduled Language of Indian Constitution. This language uses multiple script (two scripts); the first one is purely of its own origin called Meitei Mayek(Script) while another one is a borrowed Bengali Script. The very nature of resource constraint for the Meitei Script comparing to Bengali Script Manipuri compels us to think of transliteration to the output of MWE identification as another means for MWE identification in Meitei Script Manipuri. MWE plays an important role in the applications of Natural Language Processing like Machine Translation, Part of Speech tagging, Information Retrieval, Question Answering etc. Feature selection is an important factor in recognition of Manipuri MWE using CRF. This model proved to have the Recall (R) of 64.08%, Precision (P) of 86.84% and F-measure (F) of 73.74%. The transliterated output has an accuracy of 90.01% when compare with both the output of Meitei Script to Bengali Script Manipuri.**

Keywords— **MWE, Manipuri, CRF, Transliteration, Bengali Script, Meitei Mayek.**

I. Introduction

This MWE is an important topic in the application of Part of Speech Tagging, Information Retrieval, Question Answering, Summarization, Machine Translation etc. The MWE is composed of an ordered group of words which can stand independently and carries a different meaning from its constituent words. For example in English: *'to and fro', 'bye bye', 'kick the bucket' etc.* MWEs include compounds (both word-compounds and phrasal compounds), fixed expressions and technical terms. A fixed expression MWE is one whose constituent words cannot be moved randomly or substituted without distorting the overall meaning or allowing a literal interpretation. Fixed expressions range from word-compounds and collocations to idioms. Some of the proverbs and quotations can also be considered as fixed expressions.

Manipuri is a highly agglutinative Indian Language spoken mainly in Manipur. It is also spoken in some parts of Bangladesh and Myanmar. This language is a Tibeto-Burman type of language.

Manipuri uses multiple scripts that is two scripts; one is the borrowed Bengali Script while the other one is its original Meitei Mayek (Script). The development of an automatic MWE system requires either a comprehensive set of linguistically motivated rules or a large amount of annotated corpora in order to achieve reasonable performance.

The collection of corpora for the Meitei Script Manipuri is a hard task so implementing any of the MWE model is also a hard task. For those who prefer Meitei Script Manipuri we have come up with a hybrid model of MWE using CRF as in [1] and transliteration from Bengali Script and Meitei Script as in [2]. The output of MWE using CRF is a Bengali Script Manipuri which is the input to a transliteration model from Bengali Script to Meitei Script.

The feature selection of CRF model is not an easy task. The features are listed and started different combinations as features to run the CRF.

The paper is organized with related works of MWE in Manipuri and other languages are discussed in Section II which is followed by the motivation and challenges of Manipuri in Section III, Section IV is about Concepts of CRF, Section V mentions about a simple stemming rule for Manipuri, Section VI discuss about the transliteration algorithm and model used, Section VII list all the features for running the CRF, Section VIII talks about the hybrid model of MWE Manipuri, Section IX is about the experiment and the evaluation and the last Section draws the conclusion and the future works road map.

II. Related works

As far as the works of MWE are concerned then the works on MWEs can be seen in [3]-[5]. For Indian languages also works are being done to identify the MWEs [6]-[10]. The published works on identifications of NER and MWEs in Manipuri are also found. For the NER works are found in [11]-[13]. The works of MWE can be found in [1] and reduplicated MWEs in [14]-[15]. The identification of MWEs Manipuri is quite difficult since the root words in Manipuri are only noun and verb, rest of the POS are derived from them.

The stemming works of Manipuri are reported in [16]-[17]. The transliterations of Manipuri are also reported [2], [18]-[19].

III. Challenges and Motivation

A. Challenges of Indian Languages

The notable work of [11] gives us the idea about the challenges and difficulties in working with Manipuri like the other Indian Languages:

1. Unlike English and most of the European languages, Manipuri lacks capitalization information, which plays a very important role in identifying Name Entities (NEs). So it is a problem in the identification of Named Entities which may be MWEs.
2. A lot of NEs in Manipuri can appear in the dictionary with some other specific meanings. So sometimes it creates confusion between MWE NE and normal words.

3. Manipuri is a highly inflectional language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex wordforms.
4. Manipuri is a relatively free word order language. Thus NEs can appear in subject and object positions making the identification of MWE NEs task more difficult.
5. Manipuri is a resource-constrained language. Annotated corpus, name dictionaries, sophisticated morphological analyzers, POS taggers etc. are not yet available.

With the above mention challenges one need to carefully adopt a method so that optimal output is generated. These challenges also motivate us in the identification of MWE.

B. *The Agglutinative Nature*

The most important and challenging thing about Manipuri is the word structure of highly agglutinative. The affixes are bundle up one after another, specially the suffixes. Altogether 72 (seventy two) affixes are listed in Manipuri out of which 11 (eleven) are prefixes and 61 (sixty one) are suffixes. Table I shows the 10 prefixes. The prefix ম (mə) is used both as formative and pronomial prefix but it is included only once in the list. Similarly, Table II lists 55 (fifty five) suffixes as some of the suffixes are used with different forms of usage such as গুম (gum) which is used as particle as well as proposal negative, দা (də) as particle as well as locative and না (nə) as nominative, adverbial, instrumental or reciprocal.

To prove with the point that Manipuri is highly agglutinative let us site an example word: “*পুশিনহনজারমগাদাবানিদাকো*” (pusinhənjərəmgədəbənidəko), which means “(I wish I) myself would have caused to carry in (the article)”. Here there are 10 (ten) suffixes being used in a verbal root, they are “*pu*” is the verbal root which means “to carry”, “*sin*”(in or inside), “*hən*” (causative), “*jə*” (reflexive), “*rəm*” (perfective), “*gə*” (associative), “*də*” (particle), “*bə*” (infinitive), “*ni*” (copula), “*də*” (particle) and “*ko*” (endearment or wish).

TABLE I
PREFIXES

Prefixes used in Manipuri
অ, ই, ই, খু, চা, ত, থ, ন, ম and শে

TABLE II
SUFFIXES

Suffixes used in Manipuri
কন, কুম, কো, খরে, খ▯, খাই, খি, খোয়, গা, গনি, গী, গুম, ঙৈ, চা, চো, থ, থ▯, থেক, থোক, দা, দি, দুনা, দে, না, নত্তে, নি, নিং, নু, নে, পী, ফা▯, বা, বু, মক, মল, মিন, মুক, লে, লা, লক, ল্ম, লি, লী, লূ, লু, লে, লো, লোয়, শনু, শি, শিং, শিন, শু, হ▯ and হন

IV. CONCEPTS OF CONDITIONAL RANDOM FIELD (CRF)

The concept of Conditional Random Field [20] is developed in order to calculate the conditional probabilities of values on other designated input nodes of undirected graphical models. CRF encodes a conditional probability distribution with a given set of features. It is an unsupervised approach where the system learns by giving some training and can be used for testing other texts.

The conditional probability of a state sequence X=(x_1, x_2,..x_T) given an observation sequence Y=(y_1, y_2,..y_T) is calculated as :

$$P(Y|X) = \frac{1}{Z_X} \exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, X, t)\right) \quad ---(1)$$

where, $f_k(y_{t-1}, y_t, X, t)$ is a feature function whose weight λ_k is a learnt weight associated with f_k and to be learned via training. The values of the feature functions may range between -∞ … +∞, but typically they are binary. Z_X is the normalization factor:

$$Z_X = \sum_{y} \exp \sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, X, t)) \quad ---(2)$$

which is calculated in order to make the probability of all state sequences sum to 1. This is calculated as in Hidden Markov Model (HMM) and can be obtained efficiently by dynamic programming. Since CRF defines the conditional probability P(Y|X), the appropriate objective for parameter learning is to maximize the conditional likelihood of the state sequence or training data.

$$\sum_{i=1}^{N} \log P(y^i \mid x^i) \quad ---(3)$$

where, $\{(x^i, y^i)\}$ is the labeled training data.
Gaussian prior on the λ's is used to regularize the training (i.e., smoothing). If $\lambda \sim N(0,\rho^2)$, the objective function becomes,

$$\sum_{i=1}^{N} \log P(y^i \mid x^i) - \sum_{k} \frac{\lambda_i^2}{2\rho^2} \quad ---(4)$$

The objective function is concave, so the *λ's* have a unique set of optimal values.

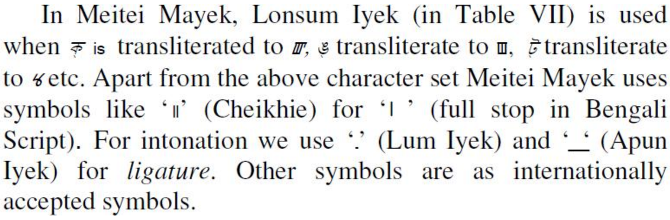
V. MANIPURI STEMMING ALGORITHM

Manipuri words are stemmed by stripping the suffixes in an iterative manner as mention in [16]. As mentioned in Section III.A, a word is rich with suffixes and prefixes. In order to stem a word an iterative method of stripping is done by using the acceptable list of prefixes (11 numbers) and suffixes (61 numbers) as mentioned in the Table I and Table II above.

VI. MANIPURI TRANSLITERATION SCHEME

The transliteration is the process of mapping a word of a source language script to another target language script. A simple transliteration scheme of Manipuri as in [20] is adopted here. Here a simple mapping of character by character from Bengali Script to Meitei Script is used. Bengali which has 52 consonants and 12 vowels is mapped to Meitei Mayek which has 27 (Twenty seven) alphabets (Iyek Ipee) and its supplements: vowels, Cheitap Iyek, Cheising Iyek and Lonsum Iyek as mention in [21] are shown in Tables III, IV, V, VI and VII.

Alphabets of Meitei Mayek are repeated uses of the same alphabet for different Bengali alphabet likeছ, শ, ষ, স in Bengali is transliterated to s in Meitei Mayek.



In Meitei Mayek, Lonsum Iyek (in Table VII) is used when ক্ is transliterated to ꯛ, ঙ্ transliterate to ꯡ, ট্ transliterate to ꯠ etc. Apart from the above character set Meitei Mayek uses symbols like ‘॥’ (Cheikhie) for ‘। ’ (full stop in Bengali Script). For intonation we use ‘.’ (Lum Iyek) and ‘_‘ (Apun Iyek) for *ligature*. Other symbols are as internationally accepted symbols.

Algorithm use for the transliteration scheme is as follows:

```
Algorithm:transliteration(line, BCC,
MMArr[], BArr[])
    line : Bengali line read from
    document
    BCC : Total number of Bengali
    Character
    MMArr[] : Bengali Characters List
    array
    BArr[] : Meitei Mayek Character
    List array
    len : Length of line
    for m = 0 to len-1 do
      tline=line.substring(m,m+1)
      if tline equals blank space
        Write a white space in the
    output file
      end of if
      else
       for index=0 to BCC-1
        if tline equals BArr[index]
         pos = index
         break
        end of if
       end of for
       Write the String MMArr[pos] in
    the output file
      end of else
    end of for
```

TABLE III
IYEK IPEE CHARACTERS IN MEITEI MAYEK

Iyek Ipee		
ক->ꯀ (kok)	স(ছ,শ,ষ)->ꯁ (Sam)	ল->ꯂ (Lai)
ম->ꯃ (Mit)	প->ꯄ (Pa)	ন->ꯅ (Na)
চ->ꯆ (Chil)	ত(ট)->ꯇ	খ->ꯈ (Khou)
ঙ-> ꯉ (Ngou)	থ(ঠ)->ꯊ(Thou)	ৱ->ꯋ (Wai)
য(য়->ꯌ (Yang)	হ->ꯍ (Huk)	উ(ঊ)->ꯎ(Un)
ই(ঈ)-ꯏ(Ee)	ফ->ꯐ (Pham)	অ->ꯑ (Atia)
গ->ꯒ (Gok)	ঝ->ꯓ(Jham)	র->ꯔ (Rai)
ব->ꯕ (Ba)	জ-> ꯖ (Jil)	দ(ড)->ꯗ(Dil)
ঘ->ꯘ (Ghou)	ধ(ঢ)->ꯙ(Dhou)	ভ->ꯚ(Bham)

TABLE IV- VOWELS OF MEITEI MAYEK

Vowel letters		
আ->ꯑꯥ(Aa)	এ->ꯑꯦ(Ae)	ঐ-ꯑꯩ(Ei)
ও->ꯑꯣ(o)	ঔ->ꯑꯧ(Ou)	অং->ꯑꯪ(Ang)

TABLE V-CHEITAP IYEK OF MEITEI MAYEK

Cheitap Iyek		
◌ো->ꯣ (ot nap)	◌ি, ◌ী-> ꯤ(inap)	◌া->ꯥ(aatap)
◌ে-> ꯦ(yetnap)	◌ৌ-> ꯧ (sounap)	◌ু, ◌ূ-> ꯨ (unap)
◌ৈ->ꯩ(cheinap)	◌ং-> ꯪ(nung)	

TABLE VI
CHEISING IYEK OR NUMERICAL FIGURES OF MEITEI MAYEK

Cheising Iyek(Numeral figure)		
১->꯱(ama)	২->꯲(ani)	৩->꯳(ahum)
৪->꯴(mari)	৫->꯵(manga)	৬->꯶(taruk)
৭->꯷(taret)	৮->꯸(nipal)	৯->꯹(mapal)
১০->꯱꯰(tara)		

TABLE VII
LONSUM IYEK OF MEITEI MAYEK

Lonsum Iyek		
ক্-> ꯛ (kok lonsum)	ল্-> ꯜ (lai lonsum)	ম্->ꯝ (mit lonsum)
প্-> ꯞ(pa lonsum)	ণ্, ন্-> ꯟ (na lonsum)	ট্,ত্->ꯠ (tillonsum)
ঙ্->ꯡ(ngou lonsum)	ই, ঈ->ꯢ(ee lonsum)	

In the transliteration scheme is diagrammatically represented in Fig. 1 and in the algorithm of *transliteration* two mapped file for Bengali Characters and corresponding Meitei Mayek Characters are used and which is used to read and stored in the *BArr* and *MMArr* arrays respectively. A test file is used so that it can compare its *index* of mapping in the Bengali Characters List file which later on used to find the corresponding target transliterated Meitei Mayek Characters Combination. The transliterated Meitei Mayek Character Combination is stored on an output file.

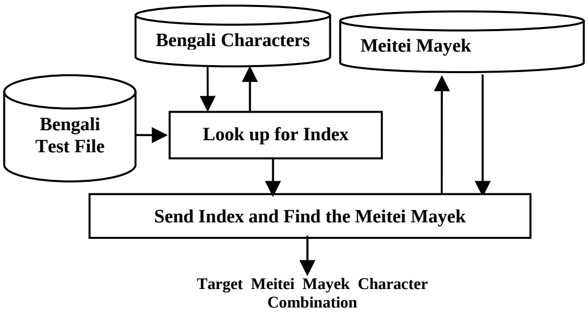

Bengali Characters
Meitei Mayek
Bengali Test File
Look up for Index
Send Index and Find the Meitei Mayek
Target Meitei Mayek Character Combination

Fig. 1 Transliteration Model of Manipuri

VII. THE INTEGRATED MODEL OF IDENTIFYING MWE MANIPURI

An integrated model of CRF based MWE identification as in [1] and transliteration in [2] is adopted but changes are made with the feature list and feature selection for running the CRF.

A. The CRF Model for Identifying MWE Manipuri

For the current work C++ based CRF++ 0.53 package[1] which is readily available as open source for segmenting or labeling sequential data is used. The CRF model for identification of Manipuri MWE (Figure 2) consists of mainly data training and data testing. The following subsection will brief about each step of this CRF model we have used.

[1] http://crfpp.sourceforge.net/

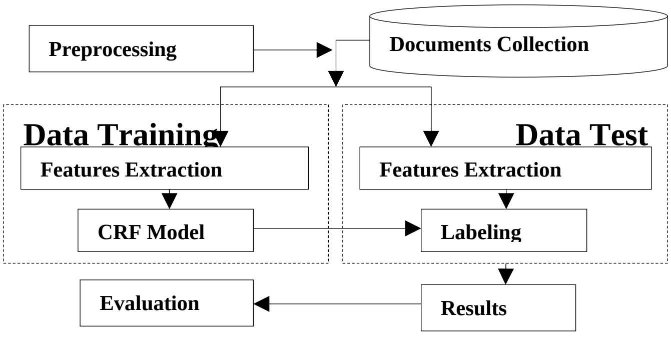


Fig. 2 CRF Model of Identifying MWE

B. The Feature List

In order to get the best result, a careful listing of features is important in CRF. The various candidate features are listed as follows:

F= {W_{i-m}, …, W_{i-1}, $W_{i,}$ $W_{i+1,\dots}$, $W_{i-n,}$ SW_{i-m}, …, SW_{i-1}, $SW_{i,}$ $SW_{i+1,\dots}$, SW_{i-n} , Acceptable suffixes present in the word, Binary notation if suffix is present, Number of acceptable suffixes, Acceptable prefixes present in the word, Binary notation if prefix is present, Binary Notation of general salutations in previous words, Binary Notation of general follow up words of Multiword Name Entities, Digit feature, Word length, Word frequency and its surrounding word frequency, Surrounding POS Tag}

The details of the set of features that have been applied for identification of MWE in Manipuri are as follows:

Current word and surrounding words: The current word and surrounding words are the focal point of MWE so selecting the current word and surrounding words as a feature is important.

Surrounding Stem words as feature: Stemming is done as mentioned in Section 5 and the preceding and following stem words of a particular word with the stem of the current word are used as features since the preceding and following words influence the present word in case of MWE.

Acceptable suffixes: 61 suffixes have been manually identified in Manipuri and the list of suffixes is used as one feature. As mentioned with an example in Section 3, suffixes are appended one after another and the maximum number of appended suffixes can be ten. So taking such cases into account, ten columns separated by space for each word to store every suffix present in the word. A "0" notation is being used in those columns when the word consists of less or no acceptable suffixes.

Acceptable prefixes as feature: 11 prefixes have been manually identified in Manipuri and the list of prefixes is used as a feature. For every word the prefix is identified and a column is created mentioning the prefix if it is present, otherwise the "0" notation is used.

Binary notation for suffix(es) present: The suffixes play an important role in Manipuri since it is a highly agglutinative language. For every word if suffix(es) is/are present during stemming a binary notation '1' is used, otherwise a '0' is stored.

Number of acceptable suffixes as feature: For every word the number of suffixes is identified during stemming, if any and the number of suffixes is used as a feature.

Binary notation for prefix(es) present: The prefixes play an important role in Manipuri since it is a highly agglutinative language. For every word if prefix(es) is/are present during stemming a binary notation '1' is used, otherwise a '0' is stored.

Binary Notation of general salutations/preceding word of Name Entity: Name Entities are generally MWEs. In order to identify the NE which are MWE, salutations like Mr., Miss, Mrs, Shri, Lt., Captain, Rs., St., Date etc that precede the Name Entity are considered as a feature for the MWE. A binary notation of '1' if used, else a '0' is used.

Binary notation of general follow up words of Name Entity: As mentioned above, Name Entities are generally MWEs. The following word of the current word can also be considered as a feature since a name may have ended up with clan name or surname or words like 'organization', 'Lup' etc for organization, words like 'Leikai', 'City' etc for places and so on. A binary notation of '1' if used else a '0' is used.

Digit features: Date, currency, weight, time etc are generally digits. Thus the digit feature is an important feature. A binary notation of '1' is used if the word consists of a digit else a '0' is used.

Length of the word: Length of the word is set to 1 if it is greater than 3. Otherwise, it is set to 0. Very short words are rarely MWE.

Word and surrounding word frequency: A range of frequencies for words in the training corpus are identified: those words with frequency <100 occurrences are set to the value 0, those words which occurs >=100 times but less than 400 times are set to 1 and so on. The word and its surrounding words frequency are considered as one feature since MWEs are rare in occurrence compared to those of determiners, conjunctions and pronouns.

Surrounding POS tag: The POS of the surrounding words are considered as an important feature since the POS of the surrounding words influence the MWE

C. Preprocessing

A Manipuri text document is used as an input file. The training and test files consist of multiple tokens. In addition, each token consists of multiple (but fixed number) columns where the columns are used by a template file. The template file gives the complete idea about the feature selection. Each token must be represented in one line, with the columns separated by white spaces (spaces or tabular characters). A sequence of tokens becomes a **sentence**. Before undergoing training and testing in the CRF, the input document is converted into a multiple token file with fixed columns and the template file allows the feature combination and selection. An example sentence formation of few words in the model for feeding in the CRF tool is shown in Fig. 3.

```
অদুগা অদু গা 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 CC O O
ইংলিস ইংলিস 0 0 0 0 0 0 0 0 0 0 0 0 ই 1 0 0 0 1 0 NNP O O
স্কুলশিং স্কুল শিং 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 NN O O
অমদি অমদি 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 5 CC O O
য়ুনিভর্সিটিশিংদা য়ুনিভর্সিটি দা শিং 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0 0 1 0 NLOC O O
লুরক্কবা লুরক্ক বা 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 VN O O
মতাঙদা মতাঙ দা 0 0 0 0 0 0 0 0 0 1 1 ম 1 0 0 0 1 1 RB O O
মমালোন্দা মমালোন্ দা 0 0 0 0 0 0 0 0 0 1 1 ম 1 0 0 0 1 0 NN O O
তাক্করিবা তাক্করি বা 0 0 0 0 0 0 0 0 0 1 1 ত 1 0 0 0 1 0 VN O O
অসিনা অ না সি 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0 1 3 PR O O
```

Fig. 3. Example Sample Sentence in the Training and the Testing File

An example of the template file which consists of feature details for two example stem words before the word, two stem words after the word, current word, the suffixes (upto a maximum of 10 suffixes), binary notation if suffix is present, number of suffixes, the prefix, binary notation of prefix is present, binary notation if digit is present, binary notation if general list of salutation or preceding word is present, binary notation if general list of follow up word is present, frequency of the word, word length, POS of the current word, POS of the prior two word, POS of the following two word details is shown in Fig. 4.

```
    # Unigram
    U00:%x[-2,1]
    U01:%x[-1,1]
    U02:%x[0,1]
    U03:%x[1,1]
     U04:%x[2,1]
    U05:%x[-1,1]
    U06:%x[0,0]
    U10:%x[0,2]
    U11:%x[0,3]
    U12:%x[0,4]
    U13:%x[0,5]
    U14:%x[0,6]
    U15:%x[0,7]
    U16:%x[0,8]
    U17:%x[0,9]
    U18:%x[0,10]
    U20:%x[0,11]
    U21:%x[0,12]
    U22:%x[0,13]
    U23:%x[0,14]
    U24:%x[0,15]
    U25:%x[0,16]
    U26:%x[0,17]
    U27:%x[0,18]
    U28:%x[0,19]
    U29:%x[0,20]
    U30:%x[0,21]
    U31:%x[-1,21]
    U32:%x[-2,21]
    U33:%x[0,21]
    U34:%x[1,21]
    U35:%x[2,21]
# Bigram
```

Fig. 4. Example template file

To run the CRF generally two standard files of multiple tokens with fixed columns are created: one for training and another one for testing. In the training file the last column is manually tagged with all those identified MWEs using the tags of B-MWE and I-MWE for the beginning and rest of the MWE respectively else 'O' for those which are not MWE. In the test file we can either use the same tagging for comparisons and evaluation or only 'O' for all the tokens regardless of whether it is MWE or not.

D. *Model File After Training*

In order to obtain a model file we train the CRF using the training file. This model file is a ready-made file by the CRF tool for use in the testing process. In other words the model file is the learnt file after the training of CRF. We do not need to use the template file and training file again since the model file consists of the detail information of the template file and training file.

E. *Testing*

The test file is the test data where sequential tags of the MWEs will be assigned else 'O' is assigned for those words which are not MWEs. This file has to be created in the same format as that of the training file, i.e., fixed number of columns with the same fields as that of training file.

The output of the testing process is a new file with an extra column which is tagged with B-MWE and I-MWE for the beginning and rest of the MWE respectively else a 'O' is tagged for those which are not MWEs

VIII. Experiment and Evaluation

Manipuri corpus are collected and filtered to rectify the spelling and syntax of a sentence by a linguist expert from Linguistic Department, Manipur University. In the corpus some words are written in English, such words are rewritten into Manipuri in order to avoid confusion or error in the output. The corpus we have collected includes 30,000 tokens which are of Gold standard.

Evaluation is done with the parameters of Recall, Precision and F-score as follows:

Recall,

$$\mathbf{R} = \frac{\textit{No of correct ans given by the system}}{\textit{No of correct ans in the text}}$$

Precision,

$$\mathbf{P} = \frac{\textit{No of ans given by the system}}{}$$

F-score,

$$\mathbf{F} = \frac{(\beta^2 + 1)\, PR}{\beta^2 P + R}$$

Where β is one, precision and recall are given equal weight.

TABLE VIIV
NOTATIONS USED IN TABLE IX

Notation	Meaning
W[-I,+J]	Words spanning from the i-th left position to the j-th right position
SW[-I,+J]	Stem words spanning from the i-th left position to the j-th right position
POS[-I, +J]	POS tags of the words spanning from the ith left to the jth right positions
BNP	Binary notation if prefix is present
BNS	Binary notation if suffix is present
NAS	Number of acceptable suffixes
PRE	Acceptable Prefix of the word
SUF[0,+J]	Acceptable Suffix of the word , where J=0,1,2,…..,10
GSP	Binary Notation of general salutations in previous words
GSF	Binary Notation of general follow up words of Mutiword Name Entities
WL	Word Length
WF[-I,+J]	Word frequency of the surrounding word
D	Digit feature

A number of problems have been faced while doing the experiment due to typical nature of the Manipuri language. In Manipuri, word category is not so distinct. The verbs are also under bound category. Another problem is to classify basic root forms according to the word class. Although the distinction between the noun class and verb classes is relatively clear; the distinction between nouns and adjectives is often vague. Distinction between a noun and an adverb becomes unclear because structurally a word may be a noun but contextually it is adverb. Further a part of root may also be a prefix, which leads to wrong tagging. The verb morphology is more complex than that of noun. Sometimes two words get fused to form a complete word

A. Experiment for selection of best feature

The experiment is performed with a total word of 30,000 words. The feature set is chosen manually with random pick up of the features.

TABLE VI
RESULTS ON THE DEVELOPMENT SET

Feature	R(in%)	P(in%)	FS(in%)
W[0,-2], SW[0,+1], POS[0,+3], SUF[0,5] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[-1,-2]	**64.08**	**86.84**	**73.74**
W[1,-2], SW[0,+2], POS[0,+3], SUF[0,5] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[-1,-2]	63.11	83.33	71.82
W[-2,+2], SW[-2,1], POS[2,+3], SUF[0,3] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[0,-3]	52.88	82.09	64.33
W[0,+2], SW[-1,1], POS[-1,+3], SUF[0,5] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[0,+2]	44.97	57.61	50.51
W[-3,-2], SW[-2,0], POS[-2,0], SUF[0,5] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[-1,+1]	38.51	63.28	47.88
W[0,-1], SW[-1,+3], POS[-1,+3], SUF[0,4] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[-2,-2]	36.38	53.82	43.41
W[-1,+4], SW[0,+3], POS[0,+3], SUF[0,5] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[-1,-2]	33.39	46.15	38.75
W[-3,-3], SW[-2,+2], POS[-2,+2], SUF[0,7] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[-2,+2]	21.53	66.51	32.53
W[-4,+4], SW[-3,+3], POS[-3,+3], SUF[0,8] , BNS, BNP, NAS, GSP, GSF, D, WL and WF[-3,-3]	19.47	74.68	30.89

The features are selected randomly with hit and trial method and experiments are performed in order to identify the best feature. The best features are those which gave the maximum recognition of MWE in a given text. In each run of the CRF tool the feature template are changed according to the feature set selected. Table no. IX shows the 10 (ten) best experimental result of identifying the MWE using CRF and Table no. VIII shows the notations used in Table no. IX. The result in terms of Recall (**R**), Precision (**P**) and F-measure (**F**). The System stops when the F-Score shows no improvement. The output that is the identified MWEs are feed to the transliteration model where the accuracy is compare between the outputs of Bengali Script to Meitei Script Manipuri

B. Evaluation and best Feature

The earlier model of CRF based identification of MWE in Manipuri as mentioned in [1] uses the following feature list with manual selection:

F= {W_{i-m}, …, W_{i-1}, $W_{i,}$ W_{i+1},… , W_{i-n} , |prefix|<=n, |suffix|<=n, Surrounding POS tag, word length, word frequency, acceptable prefix, acceptable suffix}

The list consists of surrounding words, prefixes, suffixes, surrounding POS, word length, word frequency, acceptable prefix and acceptable suffix. Improvement has been observed using reduplicated MWE as additional feature.

The model which has been adopted here has a different list and the best feature is chosen after the best performance, i.e., when saturated output is generated. The best result is considered when the system output is saturated, i.e, when there is no change in the output. This happens with the following feature:

F= {W_{i-2}, $W_{i-1,}$ W_{i}, SW_{i-1}, $SW_{i,}$ Upto 5 acceptable suffixes present in the word, Binary notation if suffix is present, Number of acceptable suffixes, Binary notation if prefix is present, Binary Notation of general salutations in previous words, Binary Notation of general follow up words of Mutiword Name Entities, Digit feature, Word length, Word frequency of previous two words, Current word POS tag, Following two words POS tag}

The best feature set in the model gives the Recall (**R**) of **64.08%**, Precision (**P**) of **86.84%** and F-measure (**F**) of **73.74%**.

The earlier model in [1] reports that the CRF based system shows a recall of 60.39%, precision of 85.53% and F-measure of 70.83%. It is also reported that with reduplicated MWEs as one feature it makes an improvement in implementation of CRF and thus the new improved recall as reported earlier is 62.24%, precision is 86.06% and F-measure is 72.24%.

The model adopted in this paper when compared with the exclusive CRF based MWE identification shows an improvement of **2.91%** in **F-Score** also when compare with the earlier model which has improvement done with reduplicated MWE shows a better **F-score** of **1.5%**.

In the case of **Recall, 3.69%** is recorded comparing with the earlier model and **1.84%** improvement is found comparing with the previous improved model with reduplicated MWE as added feature.

An improvement of **1.31%** in the **Precision** too can be noticed and **0.78%** when compared with the improved CRF using reduplicated MWE.

The accuracy in the transliteration of the identified MWEs shows and improved accuracy that is the claim [2] accuracy of **86.28%** to **90.01%**. This is because the identified MWEs are less in number so gives a calculation with **3.73%** improvement. Also the transliterated output is compare only with both the output of Meitei Script to Bengali Script Manipuri.

IX. Conclusions

So far different approaches for identification of MWE in Manipuri are found but transliteration with the output was never attempted. This model has come up with the successful implementation of transliteration of the identified MWEs in Manipuri language for the first time. This attempt is so important for this resource constrain language. This approach will be of great help for those who prefer Manipuri in Meitei Mayek. Implementation can also be tried for the other resource constrains language. This method can be tried for other Indian and other resource constrains languages with multiple scripts.

References

[1] N. Kishorjit and S. Bandyopadhyay, *Identification of MWEs Using CRF in Manipuri and Improvement Using Reduplicated MWEs*, In the Proceedings of 8th International Conference on Natural Language (ICON-2010), IIT Kharagpur, India, pp 51-57, 2010

[2] N. Kishorjit, N. Herojit Singh, Th. Sonia and B. Sivaji, *Manipuri Transliteration from Bengali Script to Meitei Mayek: A Rule Based Approach,* C. Singh et al. (Eds.):ICISIL 2011, CCIS vol..139, Part 2, pp. 195–198, Berlin, Germany: Springer-Verlag

[3] J. Enivre, and Nilson, *Multiword Units in Syntactic Parsin,* In the Proceedings of MEMURA 2004 Workshop, Lisbon, pp.39-46, 2004

[4] Koster, *Transducing Text to Multiword Unit,* In the Proceedings of MEMURA 2004 Workshop, Lisbon, pp.31-38

[5] M. T. Diab and P. Bhutada, *Verb Noun Construction MWE Token Supervised Classification,* In the Workshop on Multiword Expression. ACL-IJCNLP, Singapore , pp.17-22, 2009

[6] A. Agarwal, , B. Ray, M. Choudhury, S. Sarkar, and A. Basu, *Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenarios*, In the Proceedings of ICON 2004, Macmillan, pp. 165-174, 2004

[7] S. Dandapat, P. Mitra, and S. Sarkar, *Statistical investigation of Bengali noun-verb (N-V) collocations as multi-word-expressions*, In the Proceedings of MSPIL, Mumbai, pp 230-233, 2006

[8] A. Kunchukuttan, and O. P. Damani, *A System for Compound Nouns Multiword Expression Extraction for Hindi,* In the Proceedings of ICON 2008, Macmillan, pp. 20-29, 2008

[9] C. Tanmoy and B. Sivaji, *Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach,* In the Proceedings of COLING 2010, Multiword Expressions: from Theory to Applications (MWE 2010), Beijing, China, 2010

[10] C. Tanmoy and B. Sivaji *Identification of Noun-Noun (N-N) Collocations as Multi-Word Expressions in Bengali Corpus*, In the Processings of ICON 2010, IIT Kharagpur, India, 2010

[11] T. Doren Singh, N. Kishorjit, A. Ekbal, and B. Sivaji, *Name Entity Recognition in Manipuri Using SVM.* In the proceeding of Pacific Asia Conference on Language, Information and Computation (PACLIC 2009), Hong Kong, pp.811-818, 2009

[12] N. Kishorjit, L. Newton Singh, T. Shangkhunem, S. Bishworjit, Ng. Mayekleima Chanu and B. Sivaji, *CRF Based Name Entity Recognition (NER) in Manipuri: A Highly Agglutinative Indian language*, IEEE 10.1109/NCETACS.2011.5751390

[13] L. Newton Singh, T. Shangkhunem, N. Kishorjit, and B. Sivaji, *Name Entity Recognition (NER) in Manipuri:A Rule Based Approach,* In the Proceedings of Natioanal Conference on Computer Science and Engineering (NCCSE 2011), Imphal, India, pp.18-21, 2011

[14] N. Kishorjit, and B. Sivaji, *Identification of Reduplicated MWEs in Manipuri: A Rule based Approached*. In the Proceeding 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL-2010), San Francisco, pp 49-54, 2010

[15] N. Kishorjit, L. Dhiraj, N. Bikramjit Singh, Ng. Mayekleima Chanu, and B. Sivaji, *Identification of Reduplicated Multiword Expressions Using CRF*, A. Gelbukh (Ed.):CICLing 2011, LNCS vol.6608, Part I, pp. 41–51, Berlin, Germany: Springer-Verlag

[16] N. Kishorjit, S. Bishworjit, M. Romina, Ng. Mayekleima Chanu, and B. Sivaji, *A Light Weight Manipuri Stemmer*, In the Proceedings of Natioanal Conference on Indian Language Computing (NCILC), Chochin, India, 2010

[17] S. Bishworjit, M. Romina, N. Kishorjit and B. Sivaji, *A Transliterated Manipuri Stemmer: Bengali Script to Meitei Mayek (Script)*, In the Proceedings of Natioanal Conference on Computer Science and Engineering (NCCSE 2011), Imphal, India, pp.1-5, 2011

[18] N. Kishorjit, N. Herojit, Th. Sonia, Kh. Shinghajit, B. Sivaji, *Transliteration of Manipuri : Meitei Mayek to English Script,* In the Proceedings of International Conference on language Development and Computing Methods (ICLDCM 2010) Department of English & Department of Information Technology, Karunya University , Coimbatore,pp.240-243, 2010

[19] N. Herojit Singh, Th. Sonia, N. Kishorjit, and B. Sivaji, *Transliteration of English to Bengali Script Manipuri:A Dictionary Spell Based Approach.* In the Proceedings of Natioanal Conference on Computer Science and Engineering (NCCSE 2011), Imphal, India, pp.10-12, 2011

[20] J. Lafferty, A. McCallum, and F. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* In the Procceedings of the 18th International Conference on Machine Learning (ICML01).Williamstown, MA, USA. pp. 282-289, 2001

[21] Ng. Kangjia Mangang, *Revival of a closed account.* Sanamahi Laining Amasung Punsiron Khupham Publication, pp. 24-29, Imphal, (2003)