CHAPTER 6

MULTIWORD EXPRESSIONS DETECTION IN BENGALI

In this chapter, we describe the methods used for evaluating our systems for Automated Multiword Expression Detection with experimental results. We first conduct a series of experiments to compare and evaluate our Multiword expression extraction methods. Both human and automatic evaluations are used to measure the performance of our methods in order to understand the advantages and the problems of our approaches. Finally, we explain Multiword Expression detection elaborately.

6.1 Extracting Multiword Expressions Features

In our experiment, MWEs are extracted based on the matching of Bengali features that are given in the model as shown in the Chapter 5. . If the words are matched with system model with given features, then words are considered as MWEs as correct, otherwise system will detect them as wrong words and they will be missed.

Extraction of MWE is an important issue, where in the MWEs lexical items are attested in a predetermined corpus and extracted out into a lexicon or other lexical listing. For example, with a given verb take and preposition off, we wish to know whether the two words combine together to form a Verb Particle Constructions (VPCs) i.e. take off in a given corpus. This contrasts with MWEs identification, where the focus is on individual token instances of MWEs, although obviously extraction can be seen to be a natural consequence of identification (in compiling out the list of those attested MWEs).

It is to be assumed in MWEs extraction that there is evidence in the given corpus for each extracted MWEs to form a MWEs in some context, without making any doubt in combination of MWEs.

Matching features of Bengali Multi-word Expressions are**:**

1. If a word w has Noun as one of the possible POS tags and the word immediately preceding w (say v) also has Noun as a possible tag and has

either the possessive inflection or is uninflected, then the pair (v, w) is a possible noun- noun MWE. e.g., রাজ্য সরকার,state government

2. If w has Noun as one of the possible POS tags and v has Adjective as one of the possible tags, then the pair (v, w) is a possible adjective-noun MWE.

3. If w has Verb as one of the possible POS tags and v also has Verb as one of the possible tags and if both of them have the same inflection, but different roots, then the pair (v, w) is a possible verb-verb MWE. e.g., take a walk

4. If both v and w have Verbal Noun as a possible tag and either both has the same inflection or v has null inflection, then (v, w) is a possible verb-verb MWE.

5. Compound noun e.g., movie show, অ্যালার্ট মেসেজ
6. Compound verb(e.g., come up)
7. Conjunct Verb(e.g., effort doing)
8. Idiomatic compound Noun (মা বাবা , mother and father)
9. Reduplication (e.g.,খাওয়া দাওয়া,eating)
10. Institutionalized phrases (e.g.,শেয়ার বাজার, share market)

Implementation of all features in the model is very difficult task since works on MWEs in all languages are just growing, some of the works have been done on N-N, N-V in Bengali, we incorporate Reduplication and Idiomatic compound noun during extraction.

After the extraction, we manually validate by checking feature that were given in the model using proposed methodology in chapter 5. Therefore, the system removes those words from the list which are do not match with the given features. Only the matching MWEs will be listed as shown in our system interface result.

6.2 Algorithm for Multiword Expressions Detection

Step1: Input the Bengali text.

Step 2: Tokenize the input text.

Step 3: If the words in the form of affixation, derivation and compounding then feed the words to sentence splitter for splitting and check the words with the lexicon for matching.

Step 4: If the match is found, the words are sent to POS tagger for POS tagging.

Step 5: If the match is not found or multiple tags exist for single word then the tagger tagged the words by using statistical approach.

Step 6: Repeat step 4 and 5 till the end of the input text.

Step 7: Return the tagged output text.

Step 8: Extract new unknown new words from the tagged output as MWEs.

Step 9: Extracted tagged MWEs are put into the system model for MWEs extraction for matching given features in the model.

Step 10: Feature matching MWEs are identified using Transformation rules of N-grams.

Step 11: Repeat step 9 and step 10 till the MWEs are identified properly.

Step 12: Detect the number of MWEs present in input text as a list of MWEs.

Now we discuss the flow chart of our system model as shown in Figure.6.1

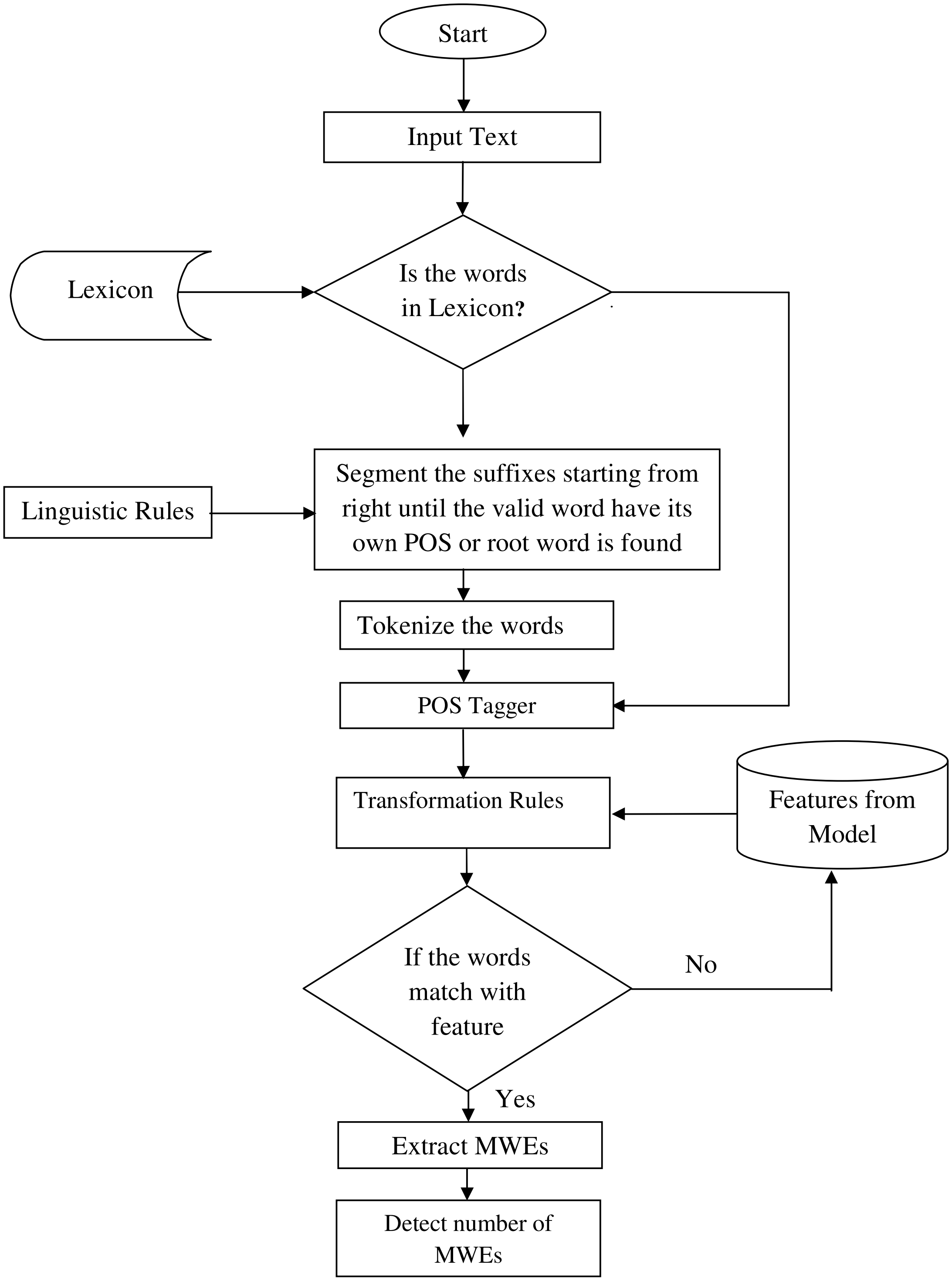


Figure 6.1 Flow chart for the system model

6.3 Steps for Proposed Flow Chart

Step 1: Enter the text.

Step 2: Check whether the text contains MWEs in the lexicon.

Step 3: If the text file containing MWEs are not found in the lexicon, then a greedy search routine walks through the word trying to find the morpheme starting from right side of the entered word for *n*-gram measure of word co-occurrences.

Step 4: Check whether the words are segmented or not

Step 5: Tokenize the words.

Step 6: Put the tokenized word in POS tagger for tagging.

Step 7: Apply transformation rules over tagged text file matching from TnT model.

Step 8: If features are matched go for extraction.

Step 9: Detect the number of MWEs**.**

6.4 Evaluation Methods

6.4.1 Measures for Multiword Eexpressions Extraction

Multiword Eexpressions are usually domain and language dependent, and for different corpora, evaluation methodologies and testing scopes are often different, leading to varied results for a given approach. Following statistical measures have been applied for our evaluation methods

6.4.1.1 Pointwise Mutual Information

The Pointwise mutual information of a pair of outcomes X and Y belonging to discrete random variables x and y compares the difference between the probability of their coincidence given their joint and marginal distribution.

Mathematically,

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x/y)}{p(x)} = \log \frac{p(y/x)}{p(y)} \qquad (6.1)$$

The mutual information (MI) of the random variables *X* and *Y* is the expected value of the PMI over all possible outcomes. The measures is symmetric PMI(x, y) = PMI(y, x) it can take positive or negative values, but it zero if x and y are independence. PMI may be negative or positive, its predictable outcome over all joint events (MI) is positive. PMI(x,y) will increase if $p(x/y)$ is fixed, but p(x) decreases.

Point wise Mutual Information (PMI) follows Chain rule as:

$$PMI(x, yz) = PMI(x, y) + PMI(x, p(z/y)) \qquad (6.2)$$

For bigram expression, it is formulated by Pecine (2005) as

$$PMI_2(x, y) = \log_2 \frac{p(x, y)}{p(x,*)p(*, y)} \qquad (6.3)$$

$$p(x, y) = \frac{f(x, y)}{N} \qquad (6.4)$$

Where P(x,y) is the Maximum Likehood(ML) estimate of the joint probability(N is the size of the corpus) and P(x,*),P(*,y)are estimation of marginal probabilities that are computed in the following way

$$p(x,*) = \frac{f(x,*)}{N} = \frac{\sum_y f(x, y)}{N} \qquad (6.5)$$

and analogically for p (*, y).

For trigrams, PMI can be calculated as follow:

$$PMI_3 = Log\frac{p(x,y,z)}{p(x,*,*)p(x,y,*)p(*,*,z)} \qquad (6.6)$$

6.4.1.2 Chi-Square Test

The Chi-Square test determines whether there is a considerable difference between the expected frequencies and the observed frequencies in one or more categories. The Chi square formula can be written as

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{i,j}} \qquad (6.7)$$

where

χ^2 is the value of Chi Square ,

$\sum$ is the sum and

O_{ij} and E_{ij} are observed and expected frequencies.

The comparison between the observed frequencies $f_{i,j}$ and the expected frequencies $E_{i,j}$ are calculated using Chi-Square method given below (Pecine, 2005)

For bigram,

$$x_2^2(x, y) = \sum \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}} \qquad (6.8)$$

Here, the expected frequency ($e_{i,j}$) and observed frequency ($f_{i,j}$) are computed by the method (Barman *et al.*,2003) given below the bigrams respectively:

$$e_{0,0} = e(x, y) = \frac{f(x,*)f(*, y)}{N} \qquad (6.9)$$

$$e_{0,1} = e(x,\neg y) = \frac{f(x,*)f(*,\neg y)}{N} \qquad (6.10)$$

And,

$$f_{0,0} = f(x, y), f_{0,1} = f(x,\neg y) = \sum_{v \neq y} f(x,v) \quad (6.11)$$

They are same for $e_{1,0}$ and $e_{1,1}$ and analogically for $f_{1,0}$ and $F_{1,1}$.

For trigram, Chi Square is computed as

$$x_3^2(x, y, z) = \sum_{i,j,k \in \{0,1\}} \frac{(f_{i,j,k} - e_{i,j,k})^2}{e_{i,j}} \qquad (6.12)$$

Here the expected frequency ($e_{i,j,k}$) and observed frequency ($f_{i,j,k}$) are computed analogically by the method formulated below for the trigram respectively:

$$e_{0,1,0} = e(x, \neg y, z) = \sum_{v \neq y} f(x, v, z) \qquad (6.13)$$

6.5 System Performance

To find out system accuracy we define association measures. Association measures are determined based on Precision (P), Recall (R) and F-measure (F) respectively. We use precision (P) and recall (R) along with a combination of the precision (P) and recall (R) to measure the comparisons between the computed results and the desirable results for Multiword Expression extraction. More specifically, we classify the results into three categories: correct, missed and wrong. When an n-gram is recognized as an MWEs by a particular approach and is also a desirable MWEs, it is considered as correct; if an n-gram is recognized as an MWEs but is not a desirable MWEs, it is considered as missed; if an n-gram is not recognized as an MWEs but is indeed a desirable MWEs, it is considered as wrong. Based on these three categories, precision (P) and recall (R) are then defined as follows:

Precision (P): Precision reflects how many of the system extracted words were correct. Precision is defined as

$$Precision(P) = \frac{correct}{(correct + wrong)}$$

Recall (R): Recall reflects how many words the system missed. Recall(R) is defined as

$$\mathrm{Re}\,call(R) = \frac{correct}{(correct + missed)}$$

Precision is a measure of accurateness, while Recall is a measure of completeness. A high recall score means that the extraction system can identify many desirable MWEs, but it does not tell how many n-grams are missed as MWEs. A high precision score means that the extraction system can identify MWEs with a high accuracy, but it may select many incorrect desirable MWEs. Often, there is an inverse relationship between recall and precision: increasing recall often decreases precision, and vice versa. In order to measure the overall performance, precision and recall need to be combined and one such composite measure is called the F measure, defined as follows:

$$F - measure(F) = \frac{2PR}{(P + R)}$$

The accuracy of the model is checked by comparing two files, one is the manually detected file and another is detected by the system based on the Bengali language features. The process of retraining of the corpus is continued till the highest level of accuracy is achieved.

Table 6.1 Experimental Result

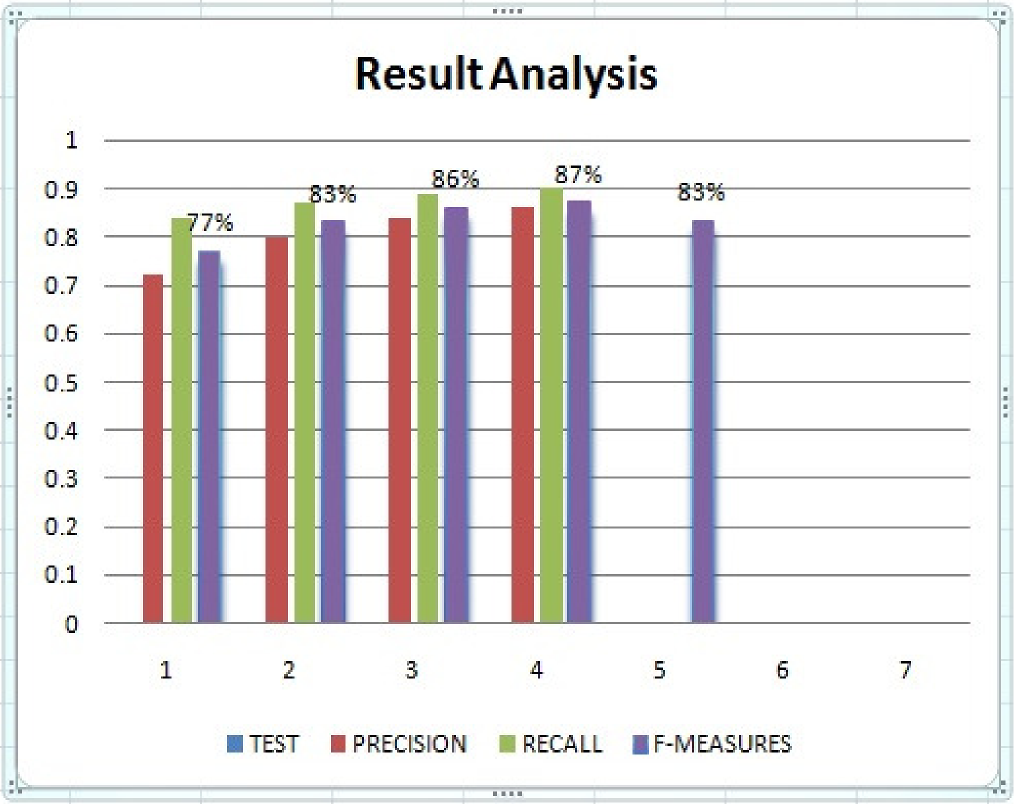
Test	No. of Words	*Hit*		No. of Words Missed	Actual Multiwords Detected
		Correct	*Wrong*		
Test 1	231	16	6	3	25
Test 2	1132	75	18	11	104
Test 3	1324	90	16	11	117
Test 4	4927	280	45	29	354

Table 6.2 Result Analysis

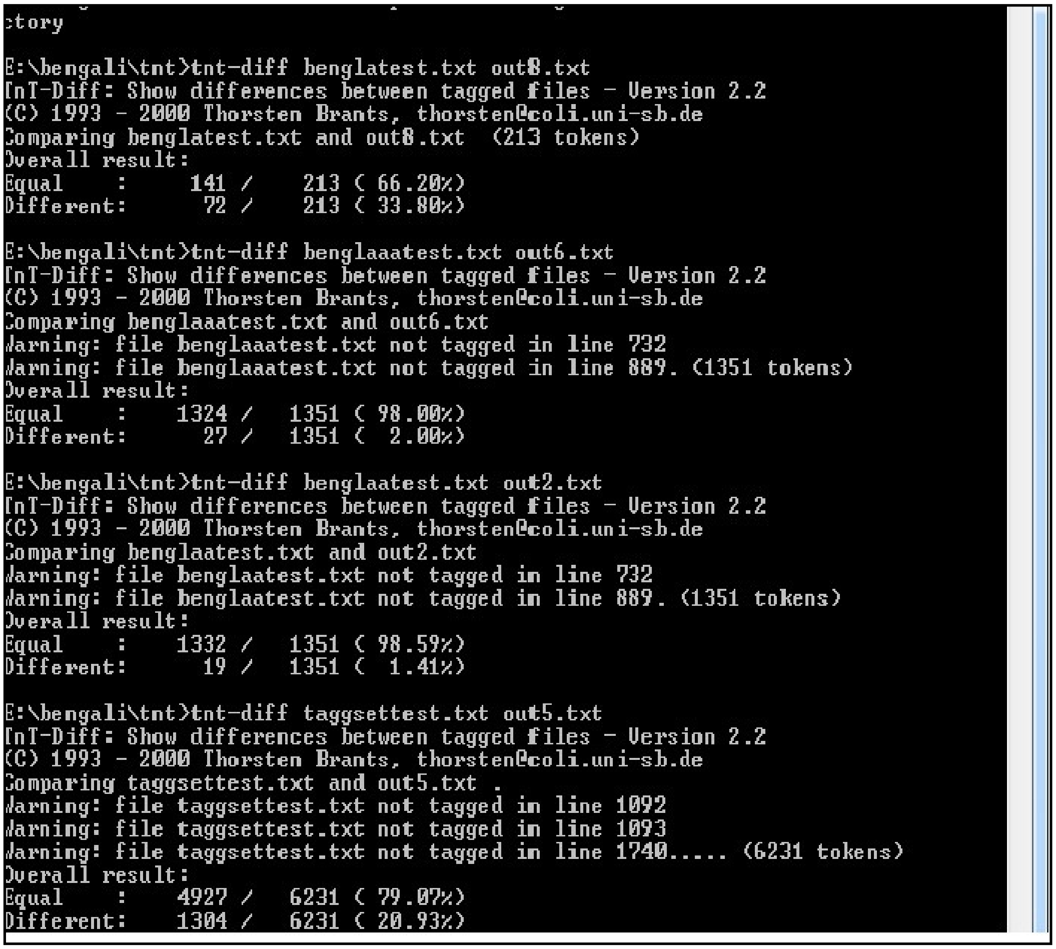
TEST	Proposed Method	Precision	Recall	F-measure
TEST1	Hybrid	72%	84%	77%
TEST2	Hybrid	80%	87%	83%
TEST3	Hybrid	84%	89%	86%
TEST4	Hybrid	86%	90%	87%
	Accuracy			83.25%

The evaluation result in our system is shown in the Table 6.2 in which maximum accuracy is found in our medium size corpus.

The evaluation result of our system analysis is shown in the table in which maximum accuracy is found in case of medium size corpus. In out experimental result, we made several tests, In test1, Accuracy of F-measure is found 77% , In test 2, accuracy of F-measure is 83% , In test 3, accuracy of F-measure is 86% and In test 4, accuracy of F-measure 87%. The overall system performance is 83.25% comparatively better than other research work.



Graph 6.2 Result Analysis



```
ctory

E:\bengali\tnt>tnt-diff benglatest.txt out8.txt
TnT-Diff: Show differences between tagged files - Version 2.2
(C) 1993 - 2000 Thorsten Brants, thorsten@coli.uni-sb.de
Comparing benglatest.txt and out8.txt  (213 tokens)
Overall result:
Equal    :     141 /    213 ( 66.20%)
Different:      72 /    213 ( 33.80%)

E:\bengali\tnt>tnt-diff benglaaatest.txt out6.txt
TnT-Diff: Show differences between tagged files - Version 2.2
(C) 1993 - 2000 Thorsten Brants, thorsten@coli.uni-sb.de
Comparing benglaaatest.txt and out6.txt
Warning: file benglaaatest.txt not tagged in line 732
Warning: file benglaaatest.txt not tagged in line 889. (1351 tokens)
Overall result:
Equal    :    1324 /   1351 ( 98.00%)
Different:      27 /   1351 (  2.00%)

E:\bengali\tnt>tnt-diff benglaatest.txt out2.txt
TnT-Diff: Show differences between tagged files - Version 2.2
(C) 1993 - 2000 Thorsten Brants, thorsten@coli.uni-sb.de
Comparing benglaatest.txt and out2.txt
Warning: file benglaatest.txt not tagged in line 732
Warning: file benglaatest.txt not tagged in line 889. (1351 tokens)
Overall result:
Equal    :    1332 /   1351 ( 98.59%)
Different:      19 /   1351 (  1.41%)

E:\bengali\tnt>tnt-diff taggsettest.txt out5.txt
TnT-Diff: Show differences between tagged files - Version 2.2
(C) 1993 - 2000 Thorsten Brants, thorsten@coli.uni-sb.de
Comparing taggsettest.txt and out5.txt .
Warning: file taggsettest.txt not tagged in line 1092
Warning: file taggsettest.txt not tagged in line 1093
Warning: file taggsettest.txt not tagged in line 1740..... (6231 tokens)
Overall result:
Equal    :    4927 /   6231 ( 79.07%)
Different:    1304 /   6231 ( 20.93%)
```

Figure 6.2 Statistical performance of system model

6.6 Multiword Expressions Detection Method

We used hybrid method for detecting MWEs. It consists of Statistical and Rule based methods. Hybrid method eliminates the weakness of individual method by improving system performance over individual methods.

Using these techniques it is possible to overcome many of the inherent limitation of single techniques, while detecting errors of any kind. We use statistical method for POS tagging and for MWEs we used rule based method.

We used rule based approach to identify MWEs which may be bigram, trigram in nature.

1. Rule based approach eliminate disambiguation that arise in MWEs due to frequent idiosyncrasies seen in MWEs.

2. The rule based system allow designing an accurate system for MWEs using linguistic feature of the language

3. Rule based approach detects number of Multiword, word form, lemma, or word sense accurately and lists them in a separate list.

4. Rule based system are comparatively easy to handle with other Indian languages

5. We used rule based methods to detect MWEs and achieved successful result.

6. Rules based method also helps us feature extraction and selection for different languages in MWEs identification that are problematic to solve.

We experiment our detection basically based on bigram, trigram words from the corpus. In case of bigram words sequence we found more accurate result up to 83.25% compared with other methods as shown in comparative analysis. After the extraction is over, we apply rule based approach for MWE detection. MWE are sorted and listed as shown in software interface.

We try to keep all the information contained in a Multiword Expression, but given our goal is to evaluate the multi-word expression extracted by auto-system which tend to be short (most of them are bi-gram), system also check for trigram but in this case system accuracy is low compared with bigram. We separate the long Multiword Expression into shorter one to get a better understanding of the performance of the system which are as follows:

a. If the expression consists of more than one independent concept/information, and each of these concepts or information consists of more than 2 words, we separate them into shorter Multiword Expressions.

   For example:

   "guide to escorts and services" → "guide to" "escorts and services"

   "মুষল ধারে বৃষ্টি হইতেছে" → "মুষল ধারে" "বৃষ্টি হইতেছে"

   If the expression consists of 2 or 3 words, we tend to keep all the words.

b. Identifying the Multiword Expressions from the corpus. System detects Multiword Expression after extraction directly. For the rest which cannot be made sure if not match with the given features.If the multi-word expression is a combination of two or more concepts system will check frequency of occurances in the corpus. For example” kick the bucket”, ‘উপজাতি’ ‘বাসস্থান’ and ঘরগৃহস্থালি etc.

c. If Multiword Expression is not a common word , using rules based approach we check the concepts consisting of more than two words separately, if they are fix collocation or regular used phrase, we keep them as multi-word expression. For example, Bengali food like ‘রস মালাই’ (Ras Malai).

d. If Multiword Expression is concept of people’s name or location or organization or specific time period. We consider them as MWEs. For examples, নতুন দিল্লী (New Delhi), আসাম বিশ্ববিদ্যালয় (Assam University) etc. Finally, based on the sense and contest of the words, words are detected. Some of the words are also missed by our system due to lack of matching features, as the given four features in our model. We thus we achieved 83.25% accuracy.

A Graphical User Interface tool named “AMEDT” has been developed by using NetBeans IDE 7.3, JDK 6 and JRE 6. The front-end of the tool has been implemented in Java and its interface is connected with a text file of Bengali lexical items called “Lexicon” as the back-end. The selection of textual database is for ease and to extend support for various platforms without the need of the installation of any DBMS server like MYSQL etc. by the end user. Each lexical item entry in “Lexicon” file has two fields:

ITEM: Bengali lexical item like any ‘text file’.
CATEGORY: The morphosyntactic category of the lexical item like NC(Common Noun), NP (Proper Noun) etc.
A screenshots view of the system tool is shown as below.

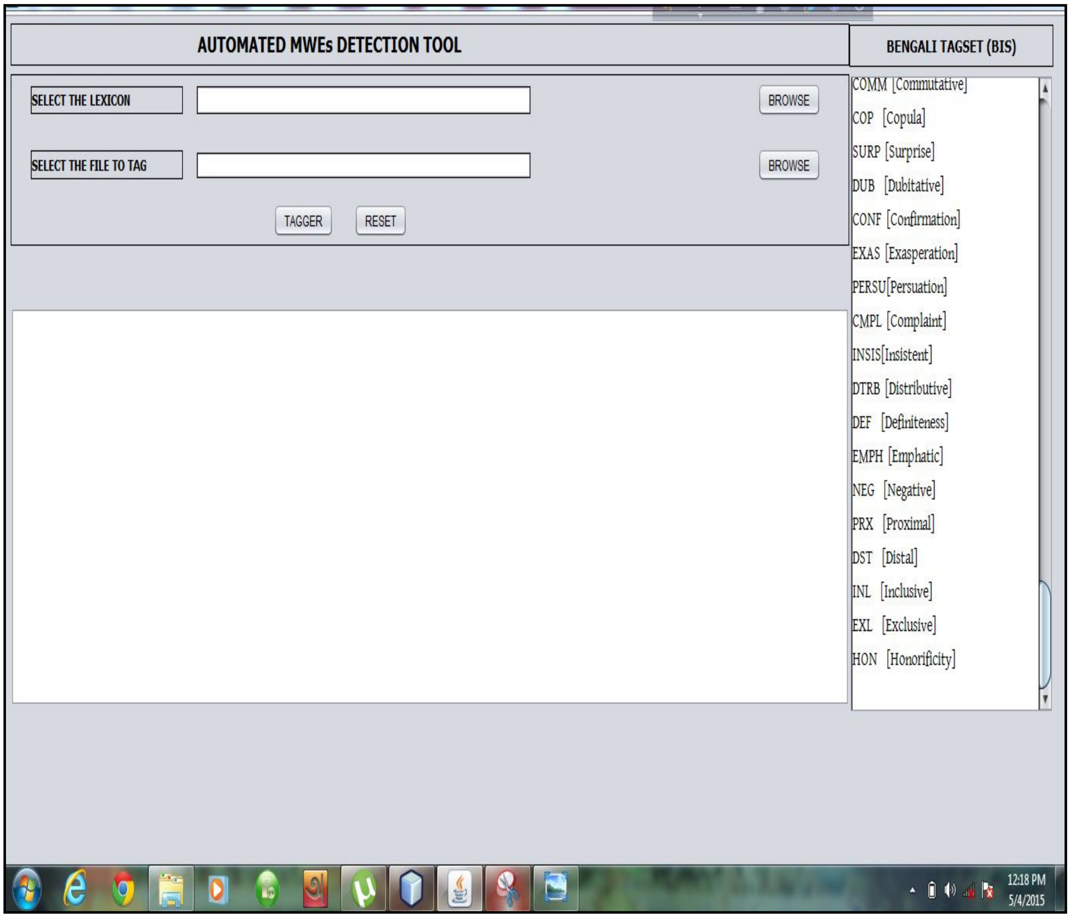
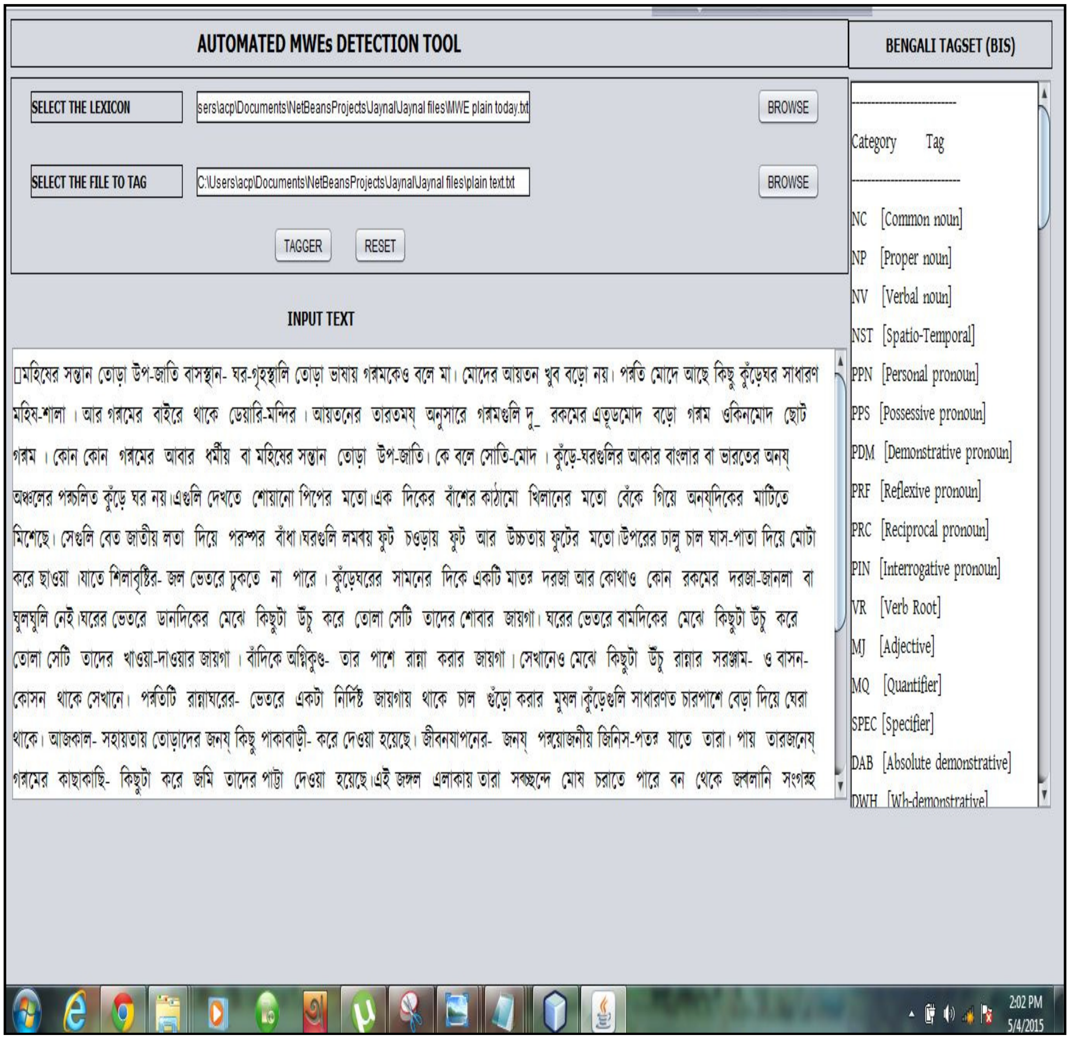


Figure 6.3 Bengali MWEs Detection Tool “AMWET” GUI Interface



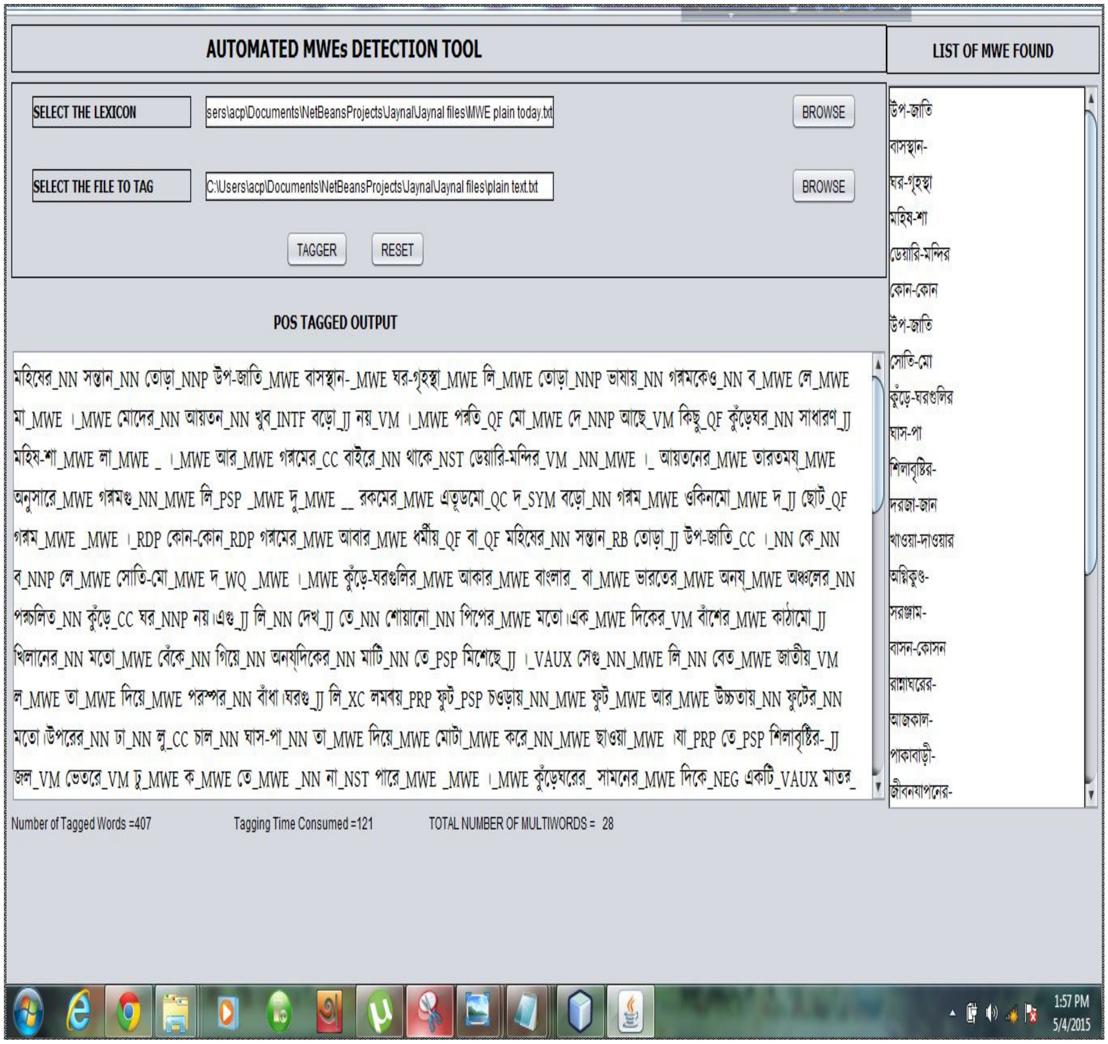
SELECT THE LEXICON: sers\acp\Documents\NetBeansProjects\Jaynal\Jaynal files\MWE plain today.txt

SELECT THE FILE TO TAG: C:\Users\acp\Documents\NetBeansProjects\Jaynal\Jaynal files\plain text.txt

□মহিষের সন্তান তোড়া উপ-জাতি বাসস্থান- ঘর-গৃহস্থালি তোড়া ভাষায় গরামকেও বলে মা। মোদের আয়তন খুব বড়ো নয়। পরতি মোদে আছে কিছু কুঁড়েঘর সাধারণ মহিষ-শালা । আর গরামের বাইরে থাকে ডেয়ারি-মন্দির । আয়তনের তারতময্ অনুসারে গরামগুলি দু_ রকমের এতুডমোদ বড়ো গরাম ওকিনমোদ ছোট গরাম । কোন কোন গরামের আবার ধর্মীয় বা মহিষের সন্তান তোড়া উপ-জাতি। কে বলে সোতি-মোদ । কুঁড়ে-ঘরগুলির আকার বাংলার বা ভারতের অনয্ অঞ্চলের পরচলিত কুঁড়ে ঘর নয়।এগুলি দেখতে শোয়ানো পিপের মতো।এক দিকের বাঁশের কাঠামো খিলানের মতো বেঁকে গিয়ে অনয্দিকের মাটিতে মিশেছে। সেগুলি বেত জাতীয় লতা দিয়ে পরস্পর বাঁধা।ঘরগুলি লমবায় ফুট চওড়ায় ফুট আর উচ্চতায় ফুটের মতো।উপরের ঢালু চাল ঘাস-পাতা দিয়ে মোটা করে ছাওয়া ।যাতে শিলাবৃষ্টির- জল ভেতরে ঢুকতে না পারে । কুঁড়েঘরের সামনের দিকে একটি মাতর দরজা আর কোথাও কোন রকমের দরজা-জানলা বা ঘুলঘুলি নেই।ঘরের ভেতরে ডানদিকের মেঝে কিছুটা উঁচু করে তোলা সেটি তাদের শোবার জায়গা। ঘরের ভেতরে বামদিকের মেঝে কিছুটা উঁচু করে তোলা সেটি তাদের খাওয়া-দাওয়ার জায়গা । বাঁদিকে অগ্নিকুণ্ড- তার পাশে রান্না করার জায়গা । সেখানেও মেঝে কিছুটা উঁচু রান্নার সরঞ্জাম- ও বাসন-কোসন থাকে সেখানে। পরতিটি রান্নাঘরের- ভেতরে একটা নির্দিষ্ট জায়গায় থাকে চাল গুঁড়ো করার মুষল।কুঁড়েগুলি সাধারণত চারপাশে বেড়া দিয়ে ঘেরা থাকে। আজকাল- সহায়তায় তোড়াদের জনয্ কিছু পাকাবাড়ী- করে দেওয়া হয়েছে। জীবনযাপনের- জনয্ পরয়োজনীয় জিনিস-পতর যাতে তারা। পায় তারজনেয্ গরামের কাছাকাছি- কিছুটা করে জমি তাদের পাট্টা দেওয়া হয়েছে।এই জঙ্গল এলাকায় তারা সবচ্ছন্দে মোষ চরাতে পারে বন থেকে জবলানি সংগরহ

Category	Tag
NC	[Common noun]
NP	[Proper noun]
NV	[Verbal noun]
NST	[Spatio-Temporal]
PPN	[Personal pronoun]
PPS	[Possessive pronoun]
PDM	[Demonstrative pronoun]
PRF	[Reflexive pronoun]
PRC	[Reciprocal pronoun]
PIN	[Interrogative pronoun]
VR	[Verb Root]
MJ	[Adjective]
MQ	[Quantifier]
SPEC	[Specifier]
DAB	[Absolute demonstrative]
DWH	[Wh-demonstrative]

Figure 6.4 Bengali MWEs Detection Tool “AMWET” GUI Interface(input text)



AUTOMATED MWEs DETECTION TOOL

LIST OF MWE FOUND

SELECT THE LEXICON: sers\acp\Documents\NetBeansProjects\Jaynal\Jaynal files\MWE plain today.txt

SELECT THE FILE TO TAG: C:\Users\acp\Documents\NetBeansProjects\Jaynal\Jaynal files\plain text.txt

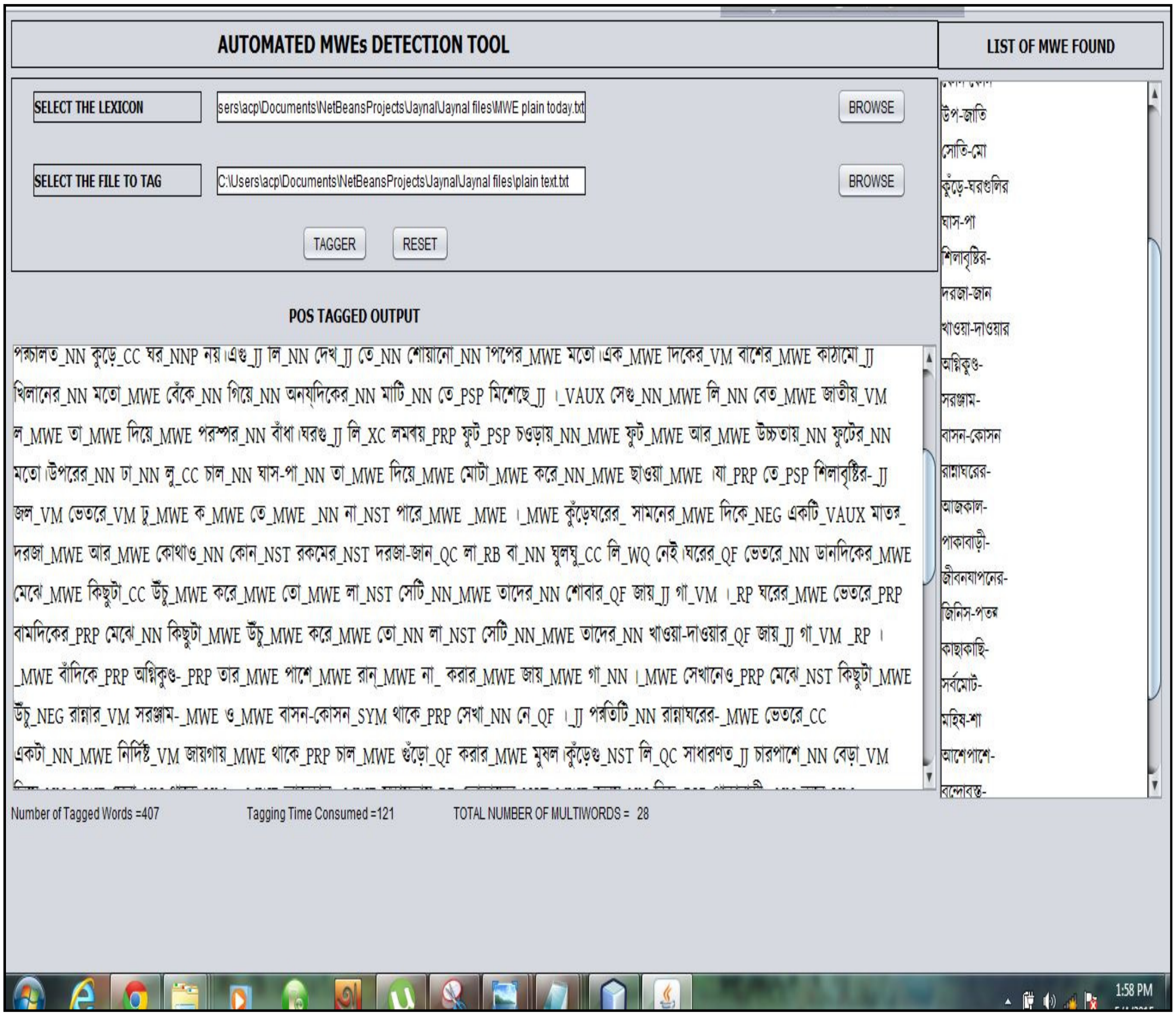
POS TAGGED OUTPUT

```
মহিষের_NN সন্তান_NN তোড়া_NNP উপ-জাতি_MWE বাসস্থান-_MWE ঘর-গৃহস্থা_MWE লি_MWE তোড়া_NNP ভাষায়_NN গরমকেও_NN ব_MWE লে_MWE
মা_MWE ।_MWE মোদের_NN আয়তন_NN খুব_INTF বড়ো_JJ নয়_VM ।_MWE পরতি_QF মো_MWE দে_NNP আছে_VM কিছু_QF কুঁড়েঘর_NN সাধারণ_JJ
মহিষ-শা_MWE লা_MWE _ ।_MWE আর_MWE গরমের_CC বাইরে_NN থাকে_NST ডেয়ারি-মন্দির_VM _NN_MWE ।_ আয়তনের_MWE তারতময্_MWE
অনুসারে_MWE গরমগু_NN_MWE লি_PSP _MWE দু_MWE __ রকমের_MWE এতূডমো_QC দ_SYM বড়ো_NN গরম_MWE ওকিনমো_MWE দ_JJ ছোট_QF
গরম_MWE _MWE ।_RDP কোন-কোন_RDP গরমের_MWE আবার_MWE ধর্মীয়_QF বা_QF মহিষের_NN সন্তান_RB তোড়া_JJ উপ-জাতি_CC ।_NN কে_NN
ব_NNP লে_MWE সোতি-মো_MWE দ_WQ _MWE ।_MWE কুঁড়ে-ঘরগুলির_MWE আকার_MWE বাংলার_ বা_MWE ভারতের_MWE অনয্_MWE অঞ্চলের_NN
পরচলিত_NN কুঁড়ে_CC ঘর_NNP নয়।এগু_JJ লি_NN দেখ_JJ তে_NN শোয়ানো_NN পিপের_MWE মতো।এক_MWE দিকের_VM বাঁশের_MWE কাঠামো_JJ
খিলানের_NN মতো_MWE বেঁকে_NN গিয়ে_NN অনয্দিকের_NN মাটি_NN তে_PSP মিশেছে_JJ ।_VAUX সেগু_NN_MWE লি_NN বেত_MWE জাতীয়_VM
ল_MWE তা_MWE দিয়ে_MWE পরস্পর_NN বাঁধা।ঘরগু_JJ লি_XC লমবায়_PRP ফুট_PSP চওড়ায়_NN_MWE ফুট_MWE আর_MWE উচ্চতায়_NN ফুটের_NN
মতো।উপরের_NN চা_NN লু_CC চাল_NN ঘাস-পা_NN তা_MWE দিয়ে_MWE মোটা_MWE করে_NN_MWE ছাওয়া_MWE ।যা_PRP তে_PSP শিলাবৃষ্টির-_JJ
জল_VM ভেতরে_VM ঢু_MWE ক_MWE তে_MWE _NN না_NST পারে_MWE _MWE ।_MWE কুঁড়েঘরের_ সামনের_MWE দিকে_NEG একটি_VAUX মাতর_
```

```
উপ-জাতি
বাসস্থান-
ঘর-গৃহস্থা
মহিষ-শা
ডেয়ারি-মন্দির
কোন-কোন
উপ-জাতি
সোতি-মো
কুঁড়ে-ঘরগুলির
ঘাস-পা
শিলাবৃষ্টির-
দরজা-জান
খাওয়া-দাওয়ার
অগ্নিকুণ্ড-
সরঞ্জাম-
বাসন-কোসন
রান্নাঘরের-
আজকাল-
পাকাবাড়ী-
জীবনযাপনের-
```

Number of Tagged Words =407   Tagging Time Consumed =121   TOTAL NUMBER OF MULTIWORDS = 28

Figure 6.5 Bengali MWEs Detection Tool "AMWET" GUI Interface(POS)



SELECT THE LEXICON: sers\acp\Documents\NetBeansProjects\Jaynal\Jaynal files\MWE plain today.txt

SELECT THE FILE TO TAG: C:\Users\acp\Documents\NetBeansProjects\Jaynal\Jaynal files\plain text.txt

```
পরচালত_NN কুড়ে_CC ঘর_NNP নয়।এগু_JJ লি_NN দেখ_JJ তে_NN শোয়ানো_NN পিপের_MWE মতো।এক_MWE দিকের_VM বাঁশের_MWE কাঠামো_JJ
খিলানের_NN মতো_MWE বেঁকে_NN গিয়ে_NN অনয্দিকের_NN মাটি_NN তে_PSP মিশেছে_JJ ।_VAUX সেগু_NN_MWE লি_NN বেত_MWE জাতীয়_VM
ল_MWE তা_MWE দিয়ে_MWE পরস্পর_NN বাঁধা।ঘরগু_JJ লি_XC লমবয়_PRP ফুট_PSP চওড়ায়_NN_MWE ফুট_MWE আর_MWE উচ্চতায়_NN ফুটের_NN
মতো।উপরের_NN ঢা_NN লু_CC চাল_NN ঘাস-পা_NN তা_MWE দিয়ে_MWE মোটা_MWE করে_NN_MWE ছাওয়া_MWE ।যা_PRP তে_PSP শিলাবৃষ্টির-_JJ
জল_VM ভেতরে_VM ঢু_MWE ক_MWE তে_MWE _NN না_NST পারে_MWE _MWE ।_MWE কুঁড়েঘরের_ সামনের_MWE দিকে_NEG একটি_VAUX মাতর_
দরজা_MWE আর_MWE কোথাও_NN কোন_NST রকমের_NST দরজা-জান_QC লা_RB বা_NN ঘুলঘু_CC লি_WQ নেই।ঘরের_QF ভেতরে_NN ডানদিকের_MWE
মেঝে_MWE কিছুটা_CC উঁচু_MWE করে_MWE তো_MWE লা_NST সেটি_NN_MWE তাদের_NN শোবার_QF জায়_JJ গা_VM ।_RP ঘরের_MWE ভেতরে_PRP
বামদিকের_PRP মেঝে_NN কিছুটা_MWE উঁচু_MWE করে_MWE তো_NN লা_NST সেটি_NN_MWE তাদের_NN খাওয়া-দাওয়ার_QF জায়_JJ গা_VM _RP ।
_MWE বাঁদিকে_PRP অগ্নিকুণ্ড-_PRP তার_MWE পাশে_MWE রান্_MWE না_ করার_MWE জায়_MWE গা_NN ।_MWE সেখানেও_PRP মেঝে_NST কিছুটা_MWE
উঁচু_NEG রান্নার_VM সরঞ্জাম-_MWE ও_MWE বাসন-কোসন_SYM থাকে_PRP সেখা_NN নে_QF ।_JJ পরতিটি_NN রান্নাঘরের-_MWE ভেতরে_CC
একটা_NN_MWE নির্দিষ্ট_VM জায়গায়_MWE থাকে_PRP চাল_MWE গুঁড়ো_QF করার_MWE মুষল।কুঁড়েগু_NST লি_QC সাধারণত_JJ চারপাশে_NN বেড়া_VM
```

Number of Tagged Words =407   Tagging Time Consumed =121   TOTAL NUMBER OF MULTIWORDS = 28

LIST OF MWE FOUND

উপ-জাতি
সোতি-মো
কুঁড়ে-ঘরগুলির
ঘাস-পা
শিলাবৃষ্টির-
দরজা-জান
খাওয়া-দাওয়ার
অগ্নিকুণ্ড-
সরঞ্জাম-
বাসন-কোসন
রান্নাঘরের-
আজকাল-
পাকাবাড়ী-
জীবনযাপনের-
জিনিস-পত্র
কাছাকাছি-
সর্বমোট-
মহিষ-শা
আশেপাশে-
বন্দোবস্ত-

Figure 6.6 Bengali MWEs Detection Tool “AMWET” GUI Interface (After Detection)

6.7 Comparative Analysis

We make comparative result analysis with different methods namely Random forest method, Single Decision Trees Method, Latent Semantic Analysis (LSA) and statistical approach as shown in the table 6.3.

Table 6.3: Comparative Analysis

Method	MWEs features selection	Accuracy
Random forest	Noun-Noun	80%
Single Decision Trees	Noun-Noun	81%
Proposed Hybrid Approach	✓ Noun-Noun ✓ Noun-Verb ✓ Reduplication ✓ Idiomatic Compound Noun	83.25% (Comparatively high with extra features)
LSA (Latent Semantic Analysis)	Noun- Noun Compound Verb particles	71.4%
Statistical Approach	Compound Noun Nominal compounds	76.61%

As shown in the table 6.3, Random forest Method used feature selection on MWEs in case of Noun-Noun only and found accuracy up to 80%, Using Single Decision trees accuracy it was found 81% in case of Noun-Noun only. Again using LSA (Latent Semantic Analysis) method accuracy was found 71.4% in case of Noun-Noun Compound and verb particles. In case of statistical approach accuracy was found 76.61% using compound noun and Nominal compound features. We used Hybrid approach on four features viz, Noun-Noun, Noun-Verb, Reduplication Idiomatic and Compound Noun and found 83.25% accuracy which is comparatively better than with other approaches as we incorporated extra features in our system model.

6.8 Chapter Summary

In this chapter, we presented experimental results of Automated Multiword Expressions detection. We find that our proposed hybrid method works well as expected. System detected Multiword Expressions contain list of MWEs which are Noun-Noun, Noun-Verb, Reduplication and Idiomatic Compound Noun with different *n*-grams. As a result we can select more meaningful MWEs and at the same time reduce the unwanted ones. To assess performance the standard information retrieval metrics are adopted are: precision, recall and F-Score. We also present comparison analysis with other methods and found comparatively better results.