Automated Multiword Expressions Detection in Bengali

Md Jaynal Abedin , Bipul Syam Purkayastha and Kh. Raju Singha

Department of Computer Science, Assam University, Silchar-788011, India
email: jaynal84@gmail.com,bipul_sh@hotmail.com, raju.singha@yahoo.com

Abstract :

Multiword Expressions (MWEs) are the combination of two or more words separated by space or delimiter which forms a new meaning instead of word individual meaning. Our works concentrate on Extraction and Automatic detection of MWEs for computational language in Bengali which are in a growing position among other Indian languages. Statistical measurement and language specific knowledge help us to extract and finally detect MWEs from medium size corpus. Natural Language processing in Bengali has started in recent years and researches in MWEs are gaining position and thus we choose to explore MWEs. Linguistics knowledge for analyzing the results has been considered and we have achieved satisfactory result.

Keywords: MWEs; NLP; Bengali; Collocation; Lexeme

Introduction:

MWEs are made up of combination of two or more than two words in which most of the time words lose their individual meaning and form a new resultant meaning. They are idiosyncratic in nature either by semantic, syntactic and lexical way. Multiword Expressions (MWEs) have been identified with an increasing amount of interest in the field of computational linguistics and Natural Language Processing (NLP) [1]. Formal definition of Multiword Expression ***is*** **[**2**]** : Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes, and (b) display lexical, syntactic, semantic, pragmatic or statistical idiomaticity. Decomposability of lexemes is that MWEs must be made up of two or more words where in most of the times the words lose their individual meanings and form a new resultant meaning. For example, *ATM Card* is potentially an MWE which is made up of two Lexemes ATM and Card, while fused word such as blue light is not considered to be a perfect MWE. However, Decomposition of an expression into multiple lexemes is still applicable. Examples of Multiword are ATM Card, Debit Card, System Error and logical Error etc. MWEs are characterized by non-compositionality, non substitutability and non-modifiability **[**3**]**. In NLP, MWEs are one of the great challenging tasks due to their high productivity and their bewildering range of syntactic, semantic, pragmatic and statistical idiomaticity they pose.

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept [3]. The basic building block of WordNet is a synset. Each synset has a unique number associated with it called as synset identity or synset id.

These synsets are created manually by lexicographers and they also manually create links between synsets representing semantic and lexical relations [4].Multiword Expressions deal with Information Retrieval (IR) System, which is an application area of computer technology for acquisition, organization, storage, retrieval, and distribution of information. Multiword detection can be applied to detect Multiword Expressions available from different repositories. Our research discipline is concerned with both the theoretical concept and the practical improvement of Multiword Expressions on reduplication Phrases, Idioms, Noun Compound (NCs), Verbal Phrases (VPs) in Bengali.

In this paper we discussed motivation of our work, various work done for extraction of MWEs in Indian languages, we also proposed system approach for MWEs extraction and to detect number of MWEs in the corpus and then the results are analyzed. Finally the conclusion is given.

MOTIVATION FOR THE IDENTIFICATION OF MWEs IN BENGALI

The fact that many difficulties arise in Bengali POS which motivated us to work on MWEs detection in Bengali. Some examples of MWEs which are difficult in POS tagging are words like কানে লাগা (kane laga) which means ‘interesting’, কান কাটা (kan kata) which means ‘shameless’, হাত থাকা (hat thaka) which means ‘right’ or’ involvement’,উঠন্ত মুলো পত্তনে চেনো যায় (uthanto mulo pottone chena jaye) which means ‘morning shows the day’, and so on. Good morphological analyzers, POS taggers, stemmer and annotated corpus etc are not yet available in this task. Bengali is highly versatile language providing one of the most challenging sets of linguistic and rich statistical features resulting in Complex and long word formation. In spite of other Natural language Processing (NLP) tasks like Information retrieval, Text summarization and Machine translation etc, in Bengali it is further needed to identify MWEs along with their extraction and detection process from different domain.

THE LITERATURE REVIEW IN MWES

The literature survey reveals the principled way to identify Multiword Expressions in different Languages. Three types of Multiword Expressions namely, Noun + Noun (compound noun), Noun + Verb (conjunct verb) and Verb + Verb (compound verb) sequences are examined. It concentrates on the linguistic methods like chunker, part-of-speech tagging, and the statistical methods like Pointwise mutual information, log-likelihood, to extract the Multiword Expressions.

The major NLP tasks relating to MWEs are: (1) identifying and extracting MWEs from corpus data and disambiguating their internal syntax and (2) interpreting MWEs. Increasingly, these tasks are being pipelined with parsers and applications such as machine translation [6]. MWEs extraction is as difficult as MWE identification in terms of syntactic flexibility and ambiguity. The bulk of research on MWE extraction has focused on extracting English verb particle constructions, light-verb constructions and idioms [7]. Based on the type of MWEs, the syntactic and semantic tasks vary based with the words combination. For example, with noun compounds, the extraction and detection tasks are relatively trivial, whereas interpretation is considerably more difficult.

OUR SYSTEM APPROACHE

We develop a system that works in offline extraction and detection MWEs which helps to extract MWEs automatically from a text corpus and then to detect the number of MWEs. This system generates a list of collocation with its rank value and the higher the rank value of the collocation in the list, the more is the probability of that collocation to fall in the category of MWEs.

Our system architecture approaches are:

Step1. Corpora collection and preprocessing

Step2. Candidate Selection

Step3. Statistical Co-occurrence tests.

Step4. Extracting MWEs

Step5.Detecting MWEs

Step1 Corpora collection and preprocessing

We have taken a row corpus of Bengali developed from a historical background. In preprocessing we need to take some special attention in various phrases like tokenization in which some words normally are not tokenized. We have taken tagged corpus as input for our system that was tagged by the POS tagger tool which is based on HMM (Hidden Markov Models).

Step2 Candidate Selection

Our algorithm selects candidate in bigram and trigram forms from sequence of the tagged Corpus. Noun, Verbs and adjective. The tagset that we used for POS tagging consist in abbreviated forms Noun Common(NC),Noun Proper(NP),Verb Main(VM) and Verb Auxiliary(VA).Thus we filter our bigram and trigram which have either NC-NC,NP-VP,NP-VA and NC-NC-NC,NC-NC-VM etc respectively.

Step 3 Statistical Co-occurrence tests

This step involves various statistical measurements through which we can test the connectedness of the collocation. It further exploits whether the pattern**s** are habitual or accidental. Also, for both the measures, frequency is counted for those same contexts containing bigrams that are either in open or hyphenated form. e.g. ধীরে ধীরে(Dheere Dheere, means slowly) ধীরে-ধীরে(Dheere Dheere, means slowly)are considered to be same in Bengali. For comparisons of the occurrence of words, we apply Pointwise Mutual Information(PMI) and to observe frequencies we used Chi-Square Test.

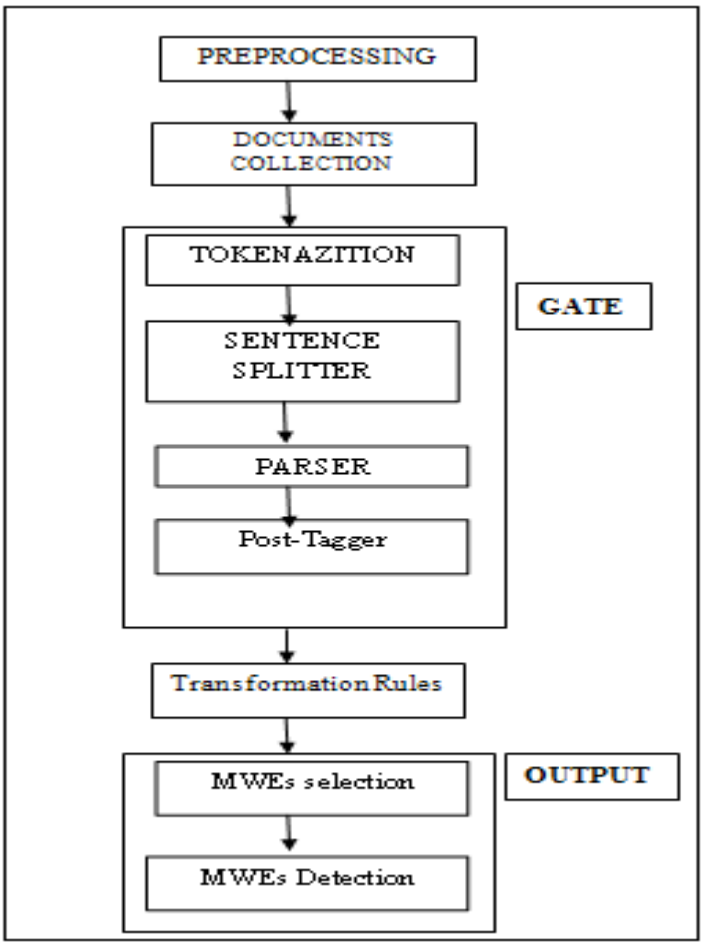

Fig.: System Architecture of MWEs Detection

Pointwise Mutual Information (PMI)

The PMI of a pair of outcomes *x* and *y* belonging to discrete random variables *X* and *Y* compares the discrepancy between the probability of their coincidence given their joint and marginal distribution. Mathematically,

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

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

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

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

For bigram expression, it is formulated [8] as

$$PMI_{2}(x,y) = \log_{2} \frac{p(x,y)}{p(x,*)\,p(*,y)}$$

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

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

$$p(x,*) = \frac{f(x,*)}{N} = \frac{\sum_{y} f(x,y)}{N}$$

And analogically for p(*,y).

For Tri-grams, PMI can be calculated as follows:

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

Chi-Square:

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

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

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 [8]

For bigram,

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

Here, the expected frequency ($e_{i,j}$) and observed frequency ($f_{i,j}$) are computed by the method given below the bigrams [9] respectively:

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

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

And,

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

They are same for e 1, 0 and e1, 1 and analogically for f1, 0 and F1, 1.

For trigram, Chi-Square is computed as

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

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:

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

Detection of Linguistically Influenced MWEs

Here we tried to find out those MWEs that are influenced by linguistic nature .Characteristics of Bengali MWEs are Repetition and Reduplication that may be complete or partial in nature. Words like Bara-Bara বর বড় (big big), Dheere Dheere ধীরে ধীরে (slowly slowly) are example of Reduplication and words like Thakur Thakur ঈশ্বর (God), Boka-soka(foolish) etc.We check whether two constituents of MWEs are synonyms of each other. We used developing Bengali WordNet for this checking purpose. We also checked if the two words are antonyms of each other. We also used syntactic and semantic analysis of the MWEs based on linguistic properties of the individual words.

Step 4 EXTRACTIONS OF MWES

Extraction MWE is also a difficult task, where, in the MWEs lexical items attested in a predetermined corpus are 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 VPC (i.e. *take off)* in a given corpus. This contrasts with MWE 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 MWE extraction that there is evidence in the given corpus for each extracted MWE to form an MWE in some context, without making any doubt in combination of MWEs.

Step 5 DETECTION OF MWES

We tried to identify those MWEs that are influenced by some linguistic phenomena like word repetition e,g. শীতশীত ভাব(seet seet bhav, feeling cold) ,Reduplication, synonyms or antonyms of the words like পাপপুণ্য(paap-punno,vice and virtue) , imitation or partial copying of the word নরম-সরম(norom-sarom, soften/softish / soft type). After the extraction, our system detects the numbers of MWEs that are present in the train corpus. We then evaluate the accuracy of the system.

RESULT ANALYSIS

The evaluation result of our system analysis is shown in the table in which maximum accuracy is found in case of medium size corpus. The differences are found between train corpus and test corpus for better system performance. In Test1, we get 98% accuracy and in Test 2, we get 85.86% accuracy. The overall system performance is better compared with other research work.

Table

The Experiment Statistics

Test	No. of Words	Overall Result	No. of MWEs Detected
Test1	1324	Equal-1324/1351(98.00%)	117
		Different-27/1351(2.00%)	
Test2	1160	Equal-1160/1351(85.86.00%)	104
		Different-191/1351(14.14.00%)	

CONCLUSION

We list out the numbers Multiword Expressions in Bengali extracted and detected in our system. It is seen that if the size of the corpus increases accuracy also increases and if the size is small, we get less accuracy. Recognition of MWEs is very complex because it varies from language to language. MWEs can be applied in real-life applications depending upon the language technologies, such as machine translation, POS etc. Study has been carried out to find optimal solution with different languages using statistical methods. Further research effort should be carried out so that analysis of MWEs can be carried in different Languages which are still in growing stage.

REFERENCES

[1] Rayson, P., Piao, S., Sharoff, S., Evert, S. & Moriron, B. V. (2010).Multiword expressions: hard going or plain sailing? Language Resources and Evaluation, vol. 44, pp. 1–5.

[2] K. Papineni, S. Roukos, T. Ward, J. Henderson and F. Reeder. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In Proceedings of Human Language Technology , San Diego, CA,pp.132-137, 2002

[3] Jennifer Brundage, M. Kresse, U. Schwall and A. Storrer.1992. Multiword lexemes:A monolingual and contrastive typology for natural language processing and machine translation. Technical Report 232, Institut fuer Wissensbasierte Systeme, IBM DeutschlandGmbH, Heidelberg.

[4] Jackendoff, Ray, The Architecture of the Language Faculty, Cambridge, MA: MIT Press, 1997.

[5] C.O.Acosta, A.Villavicenceio, P.V.Moreire, Identification and treatment of Multiword Expression applied to information Retrieval. In Proceeding of the workshop on Multiword Expression : From parsing and Generation to the real world,pages101-109,(MWE 2011).

[6] Venkatapathy, Sriram, and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In Proceedings of Coling-ACL 2006: Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, pp. 20-27.

[7] Baldwin, Timothy, and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002), 98–104, Taipei, Taiwan.

[8] P.Pecine, An extension empirical study of collocation extraction methods. ACL standard research Workshop. 2005

[9] A.K.Barman, J.Sarmah, S.K.Sarma, Automated Identification of Assameses and Bodo Multiword Expresssions. Proceeding of 2013 International conference on Advances in computing ,communication and Informatics(ICACCI) .