

Statistical Investigation of Bengali Noun-Verb (N-V) Collocations as Multi-word Expressions

Sandipan Dandapat, Pabitra Mitra, Sudeshna Sarkar

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

India 721302

{sandipan, pabitra, sudeshna}@cse.iitkgp.ernet.in

Abstract

Identification of Multi-word Expressions (MWEs) is important for several Natural Language Processing and Information Retrieval task, especially application like Machine Translation. Several statistical measurements have been developed for automatic recognition of MWEs. Recognition of MWEs requires deep or shallow syntactic pre-processing tools and large corpora. The problem is quite difficult in Bengali due to lack of such tools and large corpus. In this paper our aim is to recognize MWEs automatically from a medium size Bengali text corpus. The core of this work is an empirical study between several association measure and compute association score for each candidate N-V collocation extracted from the corpus to classify them as MWE or Non-MWE.

1 Introduction

Multi-word expressions (MWEs) are word groups whose structure and meaning (semantics) can not be derived from their component words directly, as they occur independently also. Examples include like *samaYa kATA* (meaning 'pass time'), *TikiTa kATA* (meaning 'buy ticket(s)'), *snAna karA* (meaning 'take bath'). A typical text mining system assumes each word to be a single lexical unit, but this hypothesis does not hold in case of MWEs (Becker, 1975; Fillmore 2003). They have different meaning which cross word boundaries.

A large number of MWEs have a standard syntactic structure but has different meaning. Examples of such class of collocations are

noun-verb (example '*chula kATA*' {dress hair} and '*si.N.Di bhA~ngA*' {climb the stairs}), adverb-verb (example '*musaladhAre bRRiShTi haoYA*' {rain cats and dogs}), adjective-verb (example '*lAla haoYA*' {blush}) and noun-noun (example '*mATira mAnuSha*' {down to earth}). The class of N-V collocations is important because they are frequently used.

Considerable research has been done on automatic identification of Multi-word Expressions in English, German and other European languages but not much research have been carried at this level for Bengali. Corpus frequency is used to extract plausible MWE candidates from a Bengali corpus (Agarwal et al., 2004). Various statistical association/co-occurrence measures have been suggested in literature for identification of MWEs. Some of these are Mutual Information (Church et al., 1989), Log-Likelihood (Dunning, 1993) and Saliency (Kilgarrif et al., 2001). Integration of all these statistical measure should provide better evidence for recognition of MWEs. Church and Hanks (1989) proposed a measure of association called Mutual Information. Mutual Information (MI) is the logarithm of the ratio between the probability of the two words occurring together and the probability of the each word occurring individually. The higher the MI, are more likely the words to be associated each other. The log likelihood score is the ration between the likelihood of seeing one component of a collocation given another is present and the likelihood of seeing the same component of a collocation in the absence of other. The saliency measure is an adjustment to the MI test. The saliency measure multiplies the MI score by the logarithm of the observed frequency of the word group. Thus it promotes the frequent collocations to the top ranks.

In this paper we have taken three different association measures for automatic extraction of N-V Multi-word expressions. The association measures used can be computed using only bigram collocation

tion statistics. Longer collocations are avoided as they require much larger corpus for accurate estimation of the association measures.

2 Statistical Measures Applied

Although reasonable amount of research has been done on English, German and other European language, not much work has been carried out for automatic extraction of Multi-word expressions for Bengali. Ashwini et al [2004] proposed an algorithm for automatic extraction of Multi-word expressions from a medium size Bengali raw corpus. He has used Morphological Analyzer for shallow syntactic pre-processing and statistical technique used to determine whether a collocation is a MWE or Non-MWE based on co-occurrence frequency between the words of the candidate collocations. A definition of significance of a collocation has been proposed to determine Multi-word Expressions.

The above method is based on only co-occurrence frequency where several other features have been developed for automatic extraction of MWEs. Church et al. proposed Mutual Information for measure of association. To measure the Mutual Information for a candidate collocation (xy) is defined as follows.

$$MI = \log \frac{P(xy)}{P(x)P(y)}$$

Where,

$P(xy)$ = probability of the word xy occurring together

$P(x)$ = probability of x occurring in the corpus

$P(y)$ = probability of y occurring in the corpus

These probabilities can be assigned looking at the relative bigram and unigram frequency. This method of assigning score to a candidate N-V collocation was extended by Kilgariff and Tugwell, 2001. They suggested the salience measure for better adjustment of Mutual Information. The salience measure multiplies the mutual information (MI) information score by the logarithm of the observed frequency of xy .

$$Salience = \log \frac{P(xy)}{P(x)P(y)} \cdot \log f(xy)$$

Dunnings, 1993 proposed the log-likelihood of a candidate MWE by the ratio between two likelihoods

- (i) The probability of observing one component of a collocation given the other is present.
- (ii) The probability of observing the same component of a collocation in the absence of other.

$$Log - Likelihood = -2 \sum_{i,j} f_{i,j} \log \frac{f_{i,j}}{f_{i,j}}$$

Several other measures like Z-score (Church et al., 1991) Log-Log (Kilgariff et al., 2000) etc. have been proposed.

Automatic extraction of N-V Multi-word expression can be identified as a classification task where every N-V collocation can be classified either as a MWE or as a Non-MWE. Each N-V collocation can be represented as a vector of features which are mainly several association measures. Currently three different association measures (Log-likelihood, Mutual Information and Salience) are considered to compose the feature vector. The features are extracted from a 3.7 million raw Bengali corpus. Features are selected in a way that higher the feature value, are more likely to be a MWE. For example, feature vector for N-V collocations *mAthA ghAmAno* (meaning 'brain storming') and *bhAta khAoYA* (meaning 'eat rice') are represented respectively

<salience=51.4, MI=9.4, Log-likelihood=763.9>

<salience=28.6, MI=6.7, Log-likelihood=233.1>

In the above two N-V collocations first one is a MWE and second is a Non-MWE and all the features values are much higher in case of MWE compare to Non-MWE N-V collocation. In the task we are evaluating above three association measures for automatic extraction of MWE.

3 MWE Extraction

The idea of automatic extraction of Bengali Multi-word Expressions is implemented in phases as shown in high level block diagram of Figure 1. Before the actual implementation strategy, we will discuss about the resource that has been used to implement the above idea. A large Bengali text corpus and some shallow level pre-processing tools have been used for statistical measurement to find out N-V Multi-word expressions. We have used CIIL (Central Institute of Indian Languages) Bengali corpus (around 3.7 million words) as the data for the experiment. A Bengali Part-of-Speech (POS) tagger is already developed with reasonable

accuracy (91%). The tagger assigns most probable part-of-speech tag to each word of a Bengali sentence. We have also used a Morphological Analyzer to extract the root of each word in the corpus.

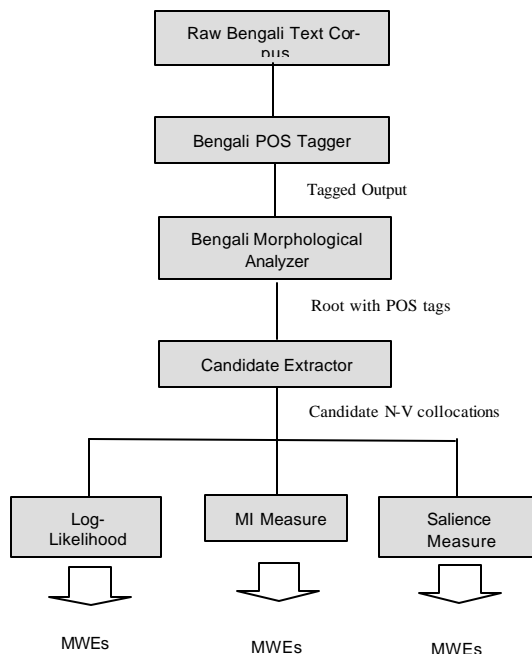


Figure 1: Block diagram for automatic extraction of N-V Multi-word Expression

3.1 Pre-processing phase

In this phase first the POS tagger is used to assign each word of the corpus with its most probable tag in the given context. In this phase we have tagged the whole raw text corpus. The tagged corpus is then passed through a Morphological Analyzer (MA) to extract the root of each word in the corpus. As Bengali is morphologically rich language so, word count does not provide much information about the frequency of a word in a text because, the word occurs with several inflections. Morphological Analyzer gives the same root for all different inflections of a particular word which provide better measurement of actual frequency of a word in text.

3.2 Candidate Item Extraction

This phase gives all possible N-V collocations that occur in a corpus. From the tagged corpus, if two consecutive words tagged as Noun and Verb respectively is extracted as a candidate N-V collocation. Currently distant N-V collocations are not considered as candidate items. Example includes like { *samaY kiChutei*

kATChe nA (samaY kATA: meaning ‘pass time’). These candidate collocations are then passed to the next phase for Log-likelihood, Mutual Information and Salience measure. As few of the candidate collocations occur with very low frequency, we are considering those candidate N-V collocations whose frequency in the corpus are greater than or equals to three. Salience measure does not require filtering of the less frequent candidate items because, the mutual information score is multiplied with the logarithmic frequency of the candidate item. However, mutual information measure needs filtering of very low frequency candidates. In order to compare the three association measures, we chose to filter out the low frequency candidate item. After preprocessing and filtering we obtained 7000 candidate N-V collocations.

3.3 Association Measures

Once we have extracted the candidate N-V collocations in the candidate item extraction phase, we have used Log-Likelihood, Mutual Information and Salience measure of each candidate N-V collocation. When the value assign by Log-Likelihood measure is large, we have more evidence to classify the candidate collocation as a MWE. Similarly, we use mutual information and salience measure to classify a candidate N-V collocation as a MWE or Non-MWE. If the Mutual Information or salience measure is large for a collocation then it’s more likely to be a Multi-word expression. Finally we sort the candidate N-V collocation on the basis of the value of log-likelihood measure and salience measure separately. Currently looking at the range of value assign to each candidate collocation by Log-likelihood, Mutual Information and Salience measure, we are extracting top 500 hundred candidate items with higher value in each measure for evaluation the accuracy.

4 Evaluation

A set of 7,000 N-V collocations are extracted from the corpus as candidate set of Multi Word Expression. The extracted candidates are sorted according to their Log-likelihood, Mutual Information and Salience measure separately. Top 500 candidates from each measure are extracted and classified into two categories (1) Valid MWE (*V*) and (2) Invalid MWE (*I*) for accuracy measure. Table 1 and shows top five candidates of the sorted list extracted by Log-likelihood measure, Mutual Information and Salience measure respectively.

Association Measure	Top-5	MWE(Y)/ (Non-MWE (N)
Log-likelihood	<i>kathA balA</i>	N
	<i>bRRidhi pAoYA</i>	Y
	<i>kathA sonA</i>	Y
	<i>hrAsa pAoYA</i>	Y
	<i>darajA kholA</i>	N
Mutual Information	<i>kutsA raTAno</i>	Y
	<i>bAdha sAdhA</i>	N
	<i>chi.DiYA bhAgA</i>	N
	<i>gatara khATAno</i>	Y
	<i>DhAka peTAno</i>	Y
Saliency	<i>darajA kholA</i>	N
	<i>mAthA ghAmAno</i>	Y
	<i>bRRidhi pAoYA</i>	Y
	<i>sigAreta dharAno</i>	Y
	<i>AghAta hAnA</i>	Y

Table 1: List of association measures, top-5 example collocations and their classification

The precision are calculated as follows.

$$Precision = \frac{V}{(V + I)}$$

The precision of Saliency, Mutual Information and Log-likelihood measure is shown in Figure 2 for first 500 candidates in the sorted list.

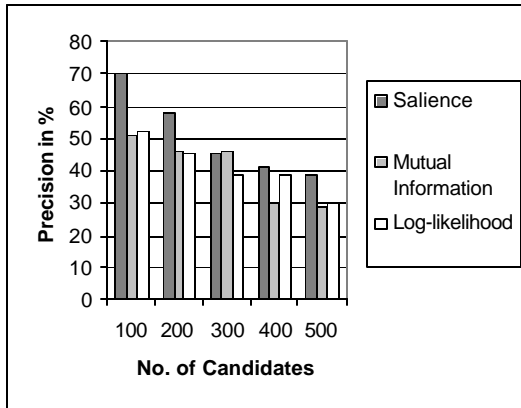


Figure 2: Precision for Bengali MWE extraction

It is quite prominent from the results of Table 1 and Figure 2 that all the three association measures are useful for automatic extraction of Bengali N-V collocation as MWEs. Saliency measure seems to be more reliable in our experimental settings as it has the highest precision for the top 500 candidates compared to Log-likelihood and Mutual Information measures.

5 Conclusion and Future Work

In this paper, we have described an approach for extraction of N-V collocation as MWE from raw corpus with the help of a POS tagger and a Morphological Analyzer using statistical association measure. We have observed three different association measures Saliency, Mutual Information and Log-likelihood. The precision achieved in Saliency measure is quite high compare to Log-likelihood and Mutual Information measure. The precision is also affected by the error propagated by the POS tagger and Morphological Analyzer.

Further development of automatic extraction of MWEs from candidate N-V collocation can focus on supervised or unsupervised learning for better discrimination of candidate N-V collocation as MWE and not MWE. More features like frequency of the candidate, Z-score and t-test etc. can be incorporated during learning.

References

- Aswini Agarwal, Biswajit Ray, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 2004. Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. In *Proceedings of ICON*, pp. 165-174, 2004.
- Joseph D. Becker. 1975. The Phrasal Lexicon. In *Theoretical Issues of NLP*, Workshop in CL, Linguistics, Psychology and AI, Cambridge, MA, pp. 60-63, 1975.
- Kenneth Wrad Church and Patrick Hans. 1990. Word Association Norms, Mutual Information and Lexicography. In *Proceedings of 27th ACL*, 16(1):22-29, 1990.
- Ted Dunning. 1993. Accurate Method for the Statistic of Surprise and Coincidence. In *Computational Linguistics*, pp. 61-74, 1993.
- Charles Fillmore. 2003. An extremist approach to multi-word expressions. A talk given at IRCS, University of Pennsylvania, 2003. www.cis.upenn.edu/~ace/kick_off_nov2003/fillmore.ppt
- Adam Kilgarrif and Joseph Rosenzweig. 2000. Framework and Results for English Senseval. *Computer and the Humanities*, 34(1):15-48, 2000.