Identification of Noun-Noun (N-N) Collocations as Multi-Word Expressions in Bengali Corpus

Tanmoy Chakraborty
Department of Computer Science and Engineering
Jadavpur University
its_tanmoy@yahoo.co.in

Abstract

Noun-Noun compounds, as a subset of Compound Nouns as well as Nominal Compounds play an important role in NLP applications like Machine Translation, Information Retrieval because of the token frequency, type frequency and their occurrence in the world’s languages. Recognition of MWEs requires deep or shallow syntactic preprocessing tools and large corpora. The problem is quite difficult in Bengali due to the lack of such tools and large corpora. This paper deals with the investigation of Noun-Noun bigram collocations from the medium-size untagged Bengali corpus of the articles of Rabindranath Tagore using simple unsupervised approach with various statistical evidences to show the affinity of the constituents of each bigram candidate as a proof of the Multi-Word Expression (MWE) and build a weighted measurement to get a distinction between MWE or non-MWE. We have mentioned different taxonomies of compound noun MWEs in Bengali based on morphosyntactic flexibility. We have also identified major Noun-Noun semantic collocations that are not MWEs. This initial approach for Bengali is promising in terms of the Precision, Recall and F-score.

1 Introduction

Over the past two decades or so, Multi-Word Expressions (MWEs; also called Multi-Word Units) have been an increasing important concern for Computational linguistics and Natural Language processing (NLP). Multi-Word expressions are those whose structure and meaning cannot be derived from their component words, as they occur independently. The term MWE has been used to refer to various types of linguistic units and expressions including idioms like ‘kick the bucket’ (‘to die’), noun compounds such as ‘village community’, phrasal verbs like ‘find out’ (‘search’) and other habitual collocations (like conjunction e.g. ‘as well as’ etc). They can be defined roughly as idiosyncratic interpretations that cross word boundaries (Sag et al., 2002). A grammatical analysis is not done while interpreting MWEs, but the entire expression is treated as a single unit.

Multi-Word Expressions are lexical items that display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity. Compositionality is the key term for MWE, which is defined to be the degree to which the features of the parts of a MWE combine to predict the feature of the whole. MWEs span a continuum from complete compositionality (also known as ‘institutionalized’ phrases e.g. ‘many thanks’ which decompose into simplex senses and generally display high syntactic variability) to partial compositionality (e.g. ‘light house’ where partial meaning is identified from the components) then to idiosyncratically compositionality (e.g. ‘spill the beans’ (to reveal) which are decomposable but coerce their parts into taking semantics unavailable outside the MWE and undergo a certain degree of syntactic variation) and finally complete non-compositionality (e.g. ‘hot dog’ where no decomposition analysis is possible and the MWE is semantically impenetrable).

1.1 Noun-Noun (N-N) Collocated MWEs

In the past few years noun compounds have received increasing attention as researchers work towards the goal of full text understanding. Compound nouns are nominal compound where two or more nouns are combined to form a single phrase such as ‘golf club’ or ‘computer science
Department' (Baldwin and Kim, 2010). There is also a broader class of nominal MWEs where the modifiers are not restricted to be nominal, but can also be verbs ('hired help') or adjectives ('open secret'). To avoid confusion, we have termed this broader set as “nominal compounds”. Compound noun MWEs can be defined as a lexical unit made up of two or more elements, each of which can function as a lexeme independent of the others(s) in other contexts and which shows some phonological and/or grammatical isolation from normal syntactic usage. One property of compound noun MWEs is their underspecified semantics. For example, while sharing the same head, there is little semantic commonality between 'nut tree', 'cloths tree' and 'family tree' (Baldwin and Kim, 2010). In each case, the meaning of the compound relates to a sense of both the head and the modifier, but the precise relationship is highly varied and not represented explicitly in any way. In English, Noun-Noun (NN) compounds occur with high frequency and high lexical and semantic variability. A summary examination of the 90 million word written component of the British National Corpus unearthed over 400,000 NN compound types, with a combined token frequency of 1.3 million; that is, over 1% of words in the BNC are NN compounds (Tanaka and Baldwin, 2003). Bengali is a language consisting of high morpho-syntactic variation at surface level. The use of compound noun multi-word expressions in Bengali is quite a common practice mainly in the literature. For example, compound nouns such as idioms like ‘taser ghar' (house of cards, fragile), institutionalized phrases like ‘ranna ghar' (kitchen), named-entities like ‘Rabindranath Thakur' (Rabindranath Tagore), numbers like ‘ek hajar asi' (one hundred eighty), kin terms like ‘pistuto bhai' (maternal cousin) etc. are very frequently used in Bengali literature.

Another common term in NLP application, which relates closely to our discussion of MWEs is collocation. A widely used definition for collocation is “an arbitrary and recurrent word combination” (Benson, 1990), or in our terms, a statistically idiomatic MWE. Collocations are often distinguished from “idioms” or “non-compositional phrases” on the grounds that these are not syntactically idiomatic and if they are semantically idiomatic, it is through a relative transparent process of figuration and metaphor.

In this work, we mainly investigate the noun-noun collocated compounds from Bengali corpus which are the subset of compound nouns and they are separated by space or hyphen. In Bengali, some compounds which are formed by two or more different words acting as a single entity are also the part of compound nouns, but morphological analysis is needed to separate their components (Dasgupta et al., 2005). So the compounds, n-grams (n>2) and named-entities are beyond the scope of our investigation. They require much larger corpus for accurate estimation of the association measures.

Reduplication is another term very frequently used in Bengali and is sometime tagged by ‘NN’. These are also not considered here as they are easy to identify because of their immediate co-occurrence and no (for complete and onomatopoetic reduplication) or minor syntactic variation in the components (for partial and correlative reduplication).

The related studies on MWEs in general and with Indian languages in particular are discussed in Section 2. Classifications of compound noun MWEs in Bengali are discussed in Section 3. The proposed system architecture and the measurement approaches are discussed in Section 4. The evaluation metrics used for evaluating the system, experimental results and the various observations we have seen in our research are discussed in Section 5 and the concluding part is drawn in Section 6.

2 Related work

A number of research activities have been carried out regarding MWE in various languages like English, German and other European languages. Various statistical co-occurrence measurements like Mutual Information (MI) (Church and Hans, 1989), Log-Likelihood (Dunning, 1993), Salience (Kilgarriff and Rosenzweig, 2001) have been suggested for identification of MWEs. In Indian languages like Hindi, a considerable approach in compound noun MWE extraction (Kunchukuttan and Damani, 2008) and a classification based approach for N-V collocations (Venkatapathy and Joshi, 2009) have been done. In Bengali, works on automated extraction of MWEs are limited in number. One method of automatic extraction of Multi-word expression in

---

1 Figuration is the property of the components of a MWE having some metaphoric, hyperbolic or metonymic in addition to their literal meaning.
Bengali (Agarwal et al. 2004) focusing mainly on Noun-Verb MWE has been carried out using significance function. In this paper we have taken five association measures like Pointwise Mutual Information (PMI), Log-likelihood ratio (LLR), Co-occurrence measure, Phi-coefficient and Significance function for automatic extraction of N-N Multi-word expressions and a combined weighted measurement technique has been proposed for final evaluation. The association measures used can be computed using only bigram collocation statistics.

The frequency of each nominal MWE is very small in a corpus. We have seen in a comparative study discussed in Section 5.2 that the results, obtained using PMI or LLR, can not identify MWEs in top ranking. So, instead of emphasizing much on frequency and its related measurements like MI, PMI, closed count, effective frequency, our system has tried to focus on probability of co-occurrence of component words in terms of their lexical affinity to each other. We have used weighted combination of features instead of Machine Learning (ML) because ML approach is language dependent and fails for narrow domain (Dias, 2003).

3 Classification of Bengali Compound Noun MWEs

As mentioned earlier, compound noun consists of more than one free morpheme and when acts as MWE, components lose their individual literal meaning and act as a single semantic unit. Compound noun MWEs can occur in open, closed or hyphenated forms and satisfy semantic non-compositionality, statistical co-occurrence or literal phenomena like reduplication etc (Kunchukuttan and Damani, 2008). Agarwal et al. (2004) have classified Bengali MWEs in three main classes consisting of twelve different fine-grained subclasses. However, taking this classification as reference and focusing on compound noun, we have classified it in seven different subclasses:

**Named-Entities (NE):** Name of the people (*Rabindranath Thakur, Rabindranath Tagore*), name of the location (*Bharat-barsa, India*), name of the organization (*Pashchim Banga Siksha Samsad, West Bengal Board of Education*) etc. where inflection can be added to the last word only.

**Idiomatic Compound Nouns:** These are unproductive and idiomatic in nature and inflection can be added only to the last word. The formation of this type is due to the hidden conjunction between the components or extinction of inflection from the first component (*maa-baba, mother and father*).

**Idioms:** They are also compound nouns with idiosyncratic meaning, but first noun is generally in possessive form (*taser ghar, fragile*). Sometimes, individual components may not carry any significant meaning and can not be a part of dictionary (*gadai laskari chal, indolent habit*). For them, no inflection is allowed even to the last word.

**Numbers:** They are highly productive, impermeable and allow slight syntactic variations like inflections. Inflections can be added only to the last component (*soya sat ghanta, seven hours and fifteen minutes*).

**Relational Noun Compounds:** They are mainly kin terms and bigram in nature. Inflection can be added with the last word (*pistuto bhai, maternal cousin*).

**Conventionalized Phrases:** Sometime they are called as ‘Institutionalized phrase’. They are not idiomatic and a particular word combination coming to be used to refer to a given object. They are productive and have unexpectedly low frequency and in doing so, contrastively highlight the statistical idiomaticity of the target expression (*bibhha barshiki, marriage anniversary*).

**Simile Terms:** They are analogy term in Bengali and sometime similar to the idioms except the fact that they are semi-productive (*hater panch, remaining resource*).

**Reduplicated Terms:** Reduplications are non-productive and tagged as noun phrase. They are further classified as onomatopoeic expressions (*khat khat, knock knock*), complete reduplication (*bara-bara, big big*), partial reduplication (*thakur-thukur, God*), semantic reduplication (*matha-mundu, head*), Correlative Reduplication (*maramari, fighting*).

A number of research activities in Bengali Named Entity detection have been carried out (Ekbal et al. 2008), but there is no such standard tool to detect this. Here we have manually identified NE. Though numbers and kin terms can be captured by some lexicons, the use of lexicons during development phase is not at all a very acceptable way. Our work mainly focuses on the extraction of productive and semi-productive bigram MWEs like idioms, idiomatic compound nouns, simile terms, numbers, relational terms, and conventionalized phrases.
4 Collocated Noun-Noun (N-N) MWE Extraction

The complete extraction procedure has been divided mainly into three phases. In the first phase, after initial pre-processing, candidate selection has been done using some heuristics to feed them into the main extraction phase. Mainly bigram collocations within same chunk have been extracted as candidates. In second phase, feature engineering consisting of various statistical co-occurrence parameters is applied on those candidates. Final decisions regarding a binary classification of MWE or non-MWE and Precision, Recall and F-score for each measurement are done in the final phase. Basic system architecture is shown in Figure 1.

4.1 Corpus Used

Resource acquisition is one of the most challenging obstacles to work with electronically resource constrained languages like Bengali. However, this system has used a large number of Bengali articles written by the noted Indian Nobel laureate Rabindranath Tagore. While we are primarily interested in single document term affinity, document information need not be maintained and manipulated by the experiment and document length normalization need not be considered. The order of the documents within the sequence is not of major importance. After merging all the articles, a medium size raw corpus has been created. It consists of 393,985 tokens and 283,533 types. Actual motivation of choosing this domain is to develop a useful statistics and further work on the Stylometry analysis.

4.2 Initial Pre-processing

The crawled corpus is so scattered and unformatted that a basic semi-automatic pre-processing has been needed. Some of them are like sentence boundary detection and make the corpus suitable for parsing. Parsing using shallow parser has been done for identifying the POS, chunk, root and inflection of each token. Some of the tokens are misspelled due to typographic or phonetic error. For example, the token ‘boi’ (book) is written as ‘বোই’ or sometime as ‘বোই’. Shallow parser is not able to detect their actual root and inflection and the number of tokens is increased. Manual identification of these redundant synonymous phonetic words is done during this phase.

4.3 Candidate Selection

After pre-processing, bigram noun sequence within the same chunk is extracted from their POS and chunk categories. Shallow parser is confused with the two noun tags i.e. common noun (‘NN’) and proper noun (‘NNP’) because of the continuous need for coinage of new terms for describing new concepts. For identifying all N-N MWEs, we have taken both of them and manual deletion of NEs has been done afterword. Although Chunking information helps to identify phrase boundary, some of the candidates belong to a chunk, which is formed by more than two nouns. Their frequency is also identified during evaluation phase. Bigram candidates can be thought of as <w1w2>. Total candidate selection phase is standing on the some heuristics described in Table 1.

Table 1. Heuristics applied in first phase.

<table>
<thead>
<tr>
<th>Heuristics</th>
<th>1. POS</th>
<th>2. Chunk</th>
<th>3. Inflection</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS of each bigram must be either ‘NN’ or ‘NNP’</td>
<td>w1 and w2 must be in the same ‘NP’ chunk</td>
<td>Inflection of w1 must be ‘null’, ‘-র (-r), ‘-এর (-er), ‘-ায় (-ay) or ‘-য়ার (-yr) and for w2, any inflection is considered</td>
<td></td>
</tr>
</tbody>
</table>

After first phase, a list of possible candidates is prepared. ‘NN’ and ‘NNP’ tags are mixed up and some of the consecutive nouns not belonging to a

---

2 http://www.rabindra-rachanabali.nltr.org

http://ltrc.iiit.ac.in/analyzer/bengali

4 Linguistic study (Chattopadhyay, 1992) reveals that for compound noun MWE, considerable inflections of first noun are only those which are mentioned above.
4.4 Statistical Feature Engineering

We have said earlier that frequency information is not a reliable source of making any statistics especially for MWE because each MWE is too low in number in a medium size corpus. We have given a proof of this assumption taking directly frequency related measures like PMI and LLR. The following are the different association measures that we have taken for our analysis:

**Pointwise Mutual Information (PMI):** The PMI of a pair of outcomes $x$ and $y$ belonging to discrete random variables quantifies the discrepancy between the probability of their coincidence given their joint distribution versus the probability of their coincidence given only their individual distributions and assuming independence (Church et al., 1990). Mathematically,

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

where, $P(xy) =$ probability of the word $x$ and $y$ 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 PMI is prone to highly overestimating the occurrence of rare events. This occurs since PMI does not incorporate the notion of support of the collocation (Kunchukuttan and Damani, 2008).

**Log-Likelihood Ratio (LLR):** The LLR is the ratio of the likelihood of the observations given the null-hypothesis to that of the alternate hypothesis (Dunning, 1993). Generally, it is the ratio between the probability of observing one component of a collocation given the other is present and the probability of observing the same component of a collocation in the absence of other.

$$\text{Log-Likelihood} = -2 \sum_{i,j} f(i,j) \log \frac{f(i,j)}{f(i) \cdot f(j)}$$

Here the order of the words in the candidate collocation was irrelevant. We have adopted first probability using Baye’s theorem by averaging the probability of $w1$ giving $w2$ and probability of $w2$ giving $w1$.

**Phi-Coefficient:** In statistics, the Phi coefficient $\Phi$ is a measure of association for two binary variables. The Phi coefficient is also related to the chi-square statistic as:

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

where $n$ is the total number of observations and $\chi^2$ is the chi-square distribution. Two binary variables are considered positively associated if most of the data falls along the diagonal cells. In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal. Here, the binary distinction denotes the positional information of the words. If we have a 2×2 table for two random variables $x$ and $y$ which denotes the presence of $w1$ and $w2$ respectively, we have following matrix:

<table>
<thead>
<tr>
<th></th>
<th>$y=1$</th>
<th>$y=0$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x=1$</td>
<td>$n_{11}$</td>
<td>$n_{10}$</td>
<td>$n_{1}$</td>
</tr>
<tr>
<td>$x=0$</td>
<td>$n_{01}$</td>
<td>$n_{00}$</td>
<td>$n_{0}$</td>
</tr>
<tr>
<td>total</td>
<td>$n_{x1}$</td>
<td>$n_{x0}$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

where, $n_{11}=$actual bigram <w1w2> count, $n_{10}=$frequency of bigram containing w1 but not w2, $n_{01}=$frequency of bigram containing w2 not w1, $n_{00}=$frequency of bigram not containing anyone of w1 and w2. $n_{x1}$ and $n_{x0}$ are the summation of their respective rows and $n_{y1}$ and $n_{y0}$ are the summation of their respective columns. Alternative words in place of absent w1 or w2 must be nouns. The phi coefficient that describes the association of x and y is

$$\phi = \frac{n_{x1}n_{y0} - n_{x0}n_{y1}}{\sqrt{n_{x1}n_{x0}n_{y1}n_{y0}}}$$

**Co-occurrence Measurement:** We have used co-occurrence measurement by using the formula adopted by Agarwal et al. (2004). It is defined as:

$$\text{co}(w1,w2) = \sum_{s \in S(w1,w2)} -d(s,w1,w2)$$

where, co(w1,w2)=co-occurrence frequency between the words (after stemming), S(w1,w2)=set of all sentences where both w1 and w2 occurs, $d(s,w1,w2)$=distance between w1 and w2 in a sentence s in terms of number of words. For every adjacent occurrence of w1 and w2, co(w1,w2) increases by 1, but if in a sentence they are largely separated, it increases only marginally. This measurement is used further in calculating significant function.

**Significance Function:** Another effective co-occurrence measurement adopted by (Agarwal et al. 2004) is used in the present work. The definition of significance function for N-N collocation is as follows:
where, \( \text{sig}_{w1}(w2) = \text{significance of } w2 \text{ with respect to } w1 \). Here slightly modification has been done from the original by interchanging the roles of \( w1 \) and \( w2 \) in the first equation and averaging them. Same modification has been done for \( f_{w1}(w2) \) which denotes number of \( w1 \) with which \( w2 \) has occurred. In the second equation, these modified values are used in their respective place. \( \text{sig}(w1,w2) \) denotes general significance of \( w1 \) and \( w2 \). \( \sigma(x) \) is the sigmoid function defined as \([\exp(-x)/(1+\exp(-x))]\). Two constants \( \kappa_1 \) and \( \kappa_2 \) define the stiffness of the sigmoid curve and for simplicity we have taken both of them as 5.0 (Agarwal et al. 2004). \( \lambda \) is defined as the average number of noun-noun co-occurrences. The value of significance function lies between 0 and 1.

**Weighted Combination**: Final evaluation has been carried out by combining all the above-mentioned features. Experimental results show that Phi-coefficient, co-occurrence and significance functions actually based on the co-occurrence distribution has given more accurate results than the frequency-based measurement approaches like LLR, PMI in the higher ranks. So these three measures are considered and have been given certain weights after working with various weights. The final results are reported for the weighted triple \(<0.45, 0.35, 0.20>\) for co-occurrence, Phi and significance function respectively. The individual scores are normalized before assigning weights so that they are in the range 0 to 1.

**System Evaluation**

**5.1 Evaluation Metrics**

We have used standard IR metrics like Precision, Recall, F-score for evaluating our final weighted measurement as well as all the association measures. Manual identification of MWEs is done for evaluation purpose. Total candidates are divided into four classes: (1) valid N-N MWEs (M), (2) valid N-N semantic collocations but not MWEs (S), (3) invalid collocations due to considering bigram in a n-gram chunk where \( n \geq 2 \) (B), (4) invalid candidates due to error in parsing like POS, chunk, inflection (E). For \( N \) number of candidates, three measuring approaches in percentage are calculated for each association measures.

Actual Precision \((V) = M/N\)

Overall Precision \((I) = (M+S)/N\)

Error rate due to B-type \((O) = B/N\)

Precision for every measure is calculated as:

**Precision** \((P) = (MWEs \text{ in top 1000 ranked candidates} / 1000)\)

Recall is defined as:

**Recall** \((R) = (MWEs \text{ in top 1000 candidates}/ \text{total N-N MWEs in the documents})\)

**F-score** \((F) = (2*P*R) / (P+R)\)

Top 1000 ranked candidates are taken to evaluate each measure in the higher ranking.

**5.2 Experimental Results**

Four classes as discussed in Section 5.1 are identified manually and their frequencies are plotted in Figure 2.

Maximum numbers of the candidates are erroneous due to parsing error. E-type candidates are filtered manually as it has produced erroneous statistics and the result might be biased. For each measurement, the scores have been sorted in descending order and the total range is divided into five ranks so that approximately equal scores fall within same rank. For every rank, three measures discussed in Section 5.1 are calculated and plotted in a graph. Table 2 depicts those results and Figure 3 gives a relative study of those measures.

![Figure 2. Frequencies of four types.](image-url)

![Figure 3. Valid N-N MWEs (in %) for each measure.](image-url)
The slope of each measure is important in this purpose because the monotonously decreasing graph indicates the more number of MWEs in upper ranks rather than in lower ranks. PMI and LLR prove to be bad measures because graphs for LLR and PMI do not follow any significant alignment and slight upward slope have been noticed. This shows the presence of higher number of MWEs in the lower ranks. Another important notification is that maximum of the lower ranked MWEs are reduplicated MWEs and they are filtered out when top 1000 ranked candidates are chosen. In weighted measured approach, maximum valid MWEs are listed in the top ranks. For this, V, I and O-measures are shown in Table 3. It is clearly shown from the column (named as V) that the corresponding graph for valid N-N MWEs is decreasing in nature. The weighted combination approach improves upon each of the individual methods. If these association measures are combined using any ranking approach, it does not require any empirical settings of weights. But the problem is that there is no standard ranking methodology on these association measures.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Weighted Measures</th>
<th>V</th>
<th>I</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>46.54</td>
<td>89.23</td>
<td>10.77</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>30.27</td>
<td>72.28</td>
<td>27.72</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>13.43</td>
<td>59.98</td>
<td>30.02</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>7.66</td>
<td>36.62</td>
<td>63.38</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>5.09</td>
<td>23.39</td>
<td>76.61</td>
</tr>
</tbody>
</table>

Table 3. Results for weighted approach.

Top three candidates for each measure and its corresponding tags are shown in Table 4. Borda’s positional ranking that does an approximate aggregation of the ranked collocation list has been used as standard ranking function in previous studies (Kunchukuttan and Damani, 2008; Dwork et al., 2001). But the results were not satisfactory using this ranking and it did not serve as an effective MWE extraction technique.

<table>
<thead>
<tr>
<th>Association Measures</th>
<th>Top 3 Candidates</th>
<th>Tags (M/S/B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI</td>
<td>ghore dhuke</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>payer dhula</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>mukher pratibimba</td>
<td>S</td>
</tr>
<tr>
<td>LLR</td>
<td>ratdin moner</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>barsar chata</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>premer sager</td>
<td>M</td>
</tr>
<tr>
<td>Co-occurrence</td>
<td>mathai hat</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>maa-r mone</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>mukher bhab</td>
<td>S</td>
</tr>
<tr>
<td>Phi-coefficient</td>
<td>khabarer kagaj</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>haater panch</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>dhanner haater</td>
<td>B</td>
</tr>
<tr>
<td>Significance</td>
<td>bayer din</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>maran dasa</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>haater panch</td>
<td>M</td>
</tr>
<tr>
<td>Weighted</td>
<td>galai dori</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>garer maath</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>nacher tale</td>
<td>S</td>
</tr>
</tbody>
</table>

Table 4. Top 3 candidates with their classifications.

Precision, Recall and F-score are performed for all association measurements as well as for the weighted approach. These are measured among top 1000 candidates after manually deleting the parsing errors. The performance metrics for different measurements are shown in Figure 4.

Figure 4. Precision, Recall and F-score (in %) for different measurements.
Precision, Recall and F-score for weighted approach are 39.64%, 91.29% and 55.28% respectively, which are quite satisfactory in the first attempt. The present work does not focus on the increase of Precision. Our goal is to make a comparative study on the existing association measurements with our own weighted measurement and try to capture maximum number of the N-N collocated MWEs with in the top 1000 ranked candidates.

As an effort of developing a lexicon on N-N MWEs has been done simultaneously, we have observed the use of MWEs by the author in the documents. For this, we have chosen 10 novels of Rabindranath Tagore randomly and made a study using the following equation:

\[ C_i = C_{i-1} \cup (CL_i - C_{i-1}) \]

where \( i \) = document id varied from 1 to 10, \( C_i = \) Combined list of N-N MWEs after \( i \)th document is processed, \( CL_i = \) Extracted list of N-N MWEs in \( i \)th document.

Value of \((CL_i - C_{i-1})\) is plotted against the document id in Figure 5, which indicates the use of new MWEs in the documents. The behavior of the graph is downward which indicates the saturation of use of MWEs with the increase of number of documents. Besides reduplications, some of the high frequent MWEs used throughout the documents are shown in Table 5 according to their frequencies.

![Figure 5. New N-N MWEs added for each document.](image)

<table>
<thead>
<tr>
<th>High frequent N-N MWEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>abalar bol</td>
</tr>
<tr>
<td>galai dori</td>
</tr>
<tr>
<td>maran dasa</td>
</tr>
<tr>
<td>khamar chokhe</td>
</tr>
<tr>
<td>khider chote</td>
</tr>
<tr>
<td>charan darshan</td>
</tr>
</tbody>
</table>

Table 5. High frequent MWEs in 10 novels.

The presence of named-entities in the candidate list also affects the performance. While conceptually all named entities are MWEs, we do not include them in our research. We have manually filtered them at the beginning of the second phase. The percentage of NE is included within the percentage of E-type error mentioned in Section 5.2.

Another important cause of taking the overall Precision (I) in consideration is that our basic goal is to build a statistics of different use of MWEs and compound in the articles by the writer and to identify the writing style or Plagiarism detection. Focusing on this, these semantically collocated compounds sometime express themselves as Institutionalized phrases in different text positions.

Significance function and co-occurrence measurement, which are used in this work, have been modified according to our need. Here, two binary variables used in Phi-coefficient are related to the positional information of constituent words w1 and w2. Weighted approach is basically a trial and error approach to find best triple.

Apart from being the first work of its kind for Bengali language, the contributions of this work are discussed as follows: (i) a morphosyntactic classification of Bengali compound noun MWE classifications beyond the conventional classification of MWEs in English (Bald-
win and Kim, 2010), (ii) new weighted approach for measuring MWEs which may be used for other types of collocation measurements, (iii) a list of Bengali N-N compound MWEs used as a lexical resource for developing synsets of MWEs, (iv) development of formatted corpus in Bengali for further study, (v) an initial study for the identification of Stylometry of Rabindranath Tagore.

6 Conclusion and Future Works

In the present work, we have identified nominal bigrams as MWEs using various statistical measurements. We have developed a list of bigram noun-noun candidates, annotated them and ranked them.

Complete identification of MWEs in Bengali is still far apart from the present work due to lack of lexical resources like WordNet. In English, two MWE types that are particularly well represented in WordNet are compound nouns (47,000 entities) and Multi-word verbs (2600 entries) (Baldwin et al., 2003). So verification using WordNet similarity is an easy approach in English. This is not possible for Bengali language. We are trying to develop such lexical resource for our purpose. Our weighted method has however given Precision of 39.64% and Recall of 91.29% at top 1000 candidates. Low Precision does not signify any bad conclusion because our main approach is to cover maximum number of MWEs in our list which has been satisfied by the high Recall. However, for our future study, we would like to apply this approach on the article of other writers and make a comparative study regarding the Stylometry of the writers. Further more, we will try to integrate Name-Entity Recognizer with the system to eliminate our manual filtering.

References


