Using Semantic Clustering for Detecting Bengali Multiword Expressions

Tanmoy Chakraborty
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur, India-721302
E-mail: its_tanmoy@cse.iitkgp.ernet.in; http://cse.iitkgp.ernet.in/ tanmoyc

Multiword Expressions (MWEs), a known nuisance for both linguistics and NLP, blur the lines between syntax and semantics. The semantic of a MWE cannot be expressed after combining the semantic of its constituents. In this study, we propose a novel approach called "semantic clustering" as an instrument for extracting the MWEs especially for resource constraint languages like Bengali. At the beginning, it tries to locate clusters of the synonymous noun tokens present in the document. These clusters in turn help measure the similarity between the constituent words of a potential candidate using a vector space model. Finally the judgment for the suitability of this phrase to be a MWE is carried out based on a predefined threshold. In this experiment, we apply the semantic clustering approach only for noun-noun bigram MWEs; however we believe that it can be extended to any types of MWEs. We compare our approach with the state-of-the-art statistical approach. The evaluation results show that the semantic clustering outperforms all other competing methods. As a byproduct of this experiment, we have started developing a standard lexicon in Bengali that serves as a productive Bengali linguistic thesaurus.

Povzetek: V prispevku je predstavljena metoda za semantično gručenje večbesednih izrazov.

1 Introduction

Over the past two decades or so, Multiword Expressions (MWEs) have been identified with an increasing amount of interest in the field of Computational linguistics and Natural Language Processing (NLP) [1]. The term "MWE" is used to refer to various types of linguistic units and expressions including idioms (kick the bucket, 'to die'), compound noun (village community), phrasal verbs (find out, 'search'), other habitual collocations like conjunctions (as well as), institutionalized phrases (many thanks) etc. However, while there is no universally agreed definition for MWE as yet, most researchers use the term to refer to those frequently occurring phrasal units which are subject to a certain level of semantic opaqueness, or non-compositionality. Sag et al. [30] defined them as "idiosyncratic interpretations that cross word boundaries (or spaces)."

MWEs are treated as a special case in semantics since individual components of an expression often fail to keep their meanings intact within the actual meaning of that expression. This opaqueness in meaning may be partial or total depending on the degree of compositionality of the whole expression [12]. MWEs have been studied for decades in Phraseology under the term "phraseological unit" [5]. But in the early 1990s, MWEs started receiving increasing attention in corpus-based computational linguistics and NLP. A number of research activities on MWEs have been carried out in various languages like English, German and many other European languages. Various statistical co-occurrence measurements like Mutual Information (MI) [15], Log-Likelihood [21], Salience [26] have been suggested for the identification of MWEs.

In the case of Indian languages, a considerable amount of research has been conducted in compound noun MWE extraction [28], complex predicate extraction [17], clustering based approach [12] and a classification based approach for identifying Noun-Verb collocations [33]. Bengali, one of the more important Indo-Iranian languages, is the sixth-most popular language in the world and spoken by a population that now exceeds 250 million[1]. Geographical Bengali-speaking population percentages are as follows: Bangladesh (over 95%), and the Indian States of Andaman & Nicobar Islands (26%), Assam (28%), Tripura (67%), and West Bengal (85%). The global total includes those which are spoken in the Diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, the United Arab Emirates, the United Kingdom, and the United States. In Bengali, works on automated extraction of MWEs are limited in number. One method of automatic extraction of Noun-Verb MWE in Bengali [2] has been carried out using morphological evidence and significance function. They have classified Bengali MWEs based on the morpho-syntactic flexibilities and proposed a statistical approach for extracting the verbal compounds from a medium size corpus.

In this paper, we propose a framework for identifying MWEs from the perspective of semantic interpretation of

---

[1]http://en.wikipedia.org/wiki/Bengali_language

MWEs that the meanings of the components are totally or partially diminished in order to construct the actual semantics of the expression. A clustering technique is employed to group all nouns that are related to the meaning of the individual component of an expression. Two types of similarity techniques based on vector space model are adapted to make a binary classification (MWE or Non-MWE) of potentially candidate phrases. We hypothesize that the more similar the components of an expression, the less probable that their combination forms a MWE. We test our hypothesis on the noun-noun bigram phrases. We also illustrate the efficiency of our model after translating the individual components of a phrase in English and fed these components into the WordNet::Similarity module module – an open-source package developed at the University of Minnesota for calculating the lexical similarity between word (or sense) pairs based on variety of similarity measure. In this paper, we test our models with different cut-off values that define the threshold of (dis)similarity and the degree of compositionality of a candidate phrase. Experimental results corroborate our hypothesis that the dissimilarity of the meaning of constituent tokens enhances the chance of constructing a MWE. The use of English WordNet, quite strikingly, substantiates its enormous productivity in identifying MWEs from Bengali documents.

The remainder of this paper is organized as follows. Section 2 introduces a preliminary study about the Bengali MWEs and their morpho-syntactic based classification. Then the detailed description of candidate selection and the baseline system are described in section 3 and section 4 respectively. Section 5 illustrates traditional statistical methodologies for extracting MWEs from the document. Section 6 presents an elaborate description of semantic clustering approach. The introduction of English WordNetSimilarity in identifying Bengali MWEs is presented in section 7. The metrics used for evaluating the systems and experimental results are discussed in section 8. The discussion regarding the utilities and shortcomings of our model is illustrated in section 9 and the concluding part is drawn in section 10.

2 Multiword expressions (MWEs)

Though MWEs are understood quite easily and their acquisition presents no difficulty to native speakers (though it is usually not the case for second language learners), it is hard to identify what features distinguish MWEs from free word combinations. Concerning this issue, the following MWE properties are mentioned in the literature: reduced syntactic and semantic transparency; reduced or lack of compositionality; more or less frozen or fixed status; possible violation of some otherwise general syntactic patterns or rules; a high degree of lexicalization (depending on pragmatic factors); a high degree of conventionality [8].

No consensus exists so far on the definition of MWEs, but almost all formulations found in research papers em-

phasize the idiosyncratic nature of this linguistic phenomenon by indicating that MWEs are "idiosyncratic interpretations that cross word boundaries (or spaces)" [30]; "a sequence of words that acts as a single unit at some level of linguistic analysis, ... they are usually instances of well productive syntactic patterns which nevertheless exhibit a peculiar lexical behavior" [8]; "a MWE is composed of two or more words that together form a single unit of meaning, e.g., frying pan, take a stroll, and kick the bucket, . . . Semantic idiosyncrasy, i.e., the overall meaning of a MWE diverges from the combined contribution of its constituent parts" [24].

2.1 Noun-Noun MWEs

In the past few years, noun compounds have been a constant source of concern to the researchers towards the goal of full text understanding [5, 7]. Compound nouns are nominal compounds where two or more nouns are combined to form a single phrase such as 'golf club' or 'computer science department' [5]. There is also a broader class of nominal MWEs where the modifiers are not restricted to be nominal, but can also be verbs (e.g., hired help) or adjectives (e.g., open secret). To avoid confusion in this article, we will use the term compound nouns when referring to this broader class, throughout the paper, we term this broader class.

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 other(s) when they occur separately in different contexts of the document. The combination of these constituents shows some phonological and/or grammatical isolation from their normal syntactic usages. One property of compound noun MWEs is their underspecified semantics. For example, while sharing the same "head noun" (i.e., rightmost noun in the noun compound), there is less semantic commonality between the components such as 'nut tree', 'cloths tree' and 'family tree' [5]. In each case, the meaning of the compound nouns 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. Noun-Noun (NN) compounds are the subset of the compound nouns consisting of two consecutive nouns side by side. In English, NN compounds occur in general with high frequency and high lexical and semantic variabilities. 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 [32].

In Bengali, similar observations are noticed when dealing with the various types of multiword expressions like compound nouns (taser ghar, 'house of cards', 'fragile'), complex predicates such as conjunct verbs (*anuvab kara*, 'to feel') and compound verbs (*uthe para*, 'to arise'), idioms (*matir manus*, 'down to the earth'), named-entities (*Rabindranath Thakur*, 'Rabindranath Tagore') etc. Ben-

gali is a language consisting of high morpho-syntactic variation at the surface level. The use of NN multiword expressions in Bengali is quite common. For example, NN compounds especially, idioms (*taser ghar*, 'fragile'), institutionalized phrases (*ranna ghar*, 'kitchen'), named-entities (*Rabindranath Thakur*, 'Rabindranath Tagore'), numbers (*panchso noi*, 'five hundred and nine'), kin terms (*pistuto bhai*, 'maternal cousin') etc. are very frequently used in Bengali literature. In the next subsection, we classify the compound nouns occurred in Bengali based on their morpho-syntactic properties.

2.2 Classifications of Bengali compound noun MWEs

Compound noun MWEs can occur in open (components are separated by space(s)), closed (components are melded together) or hyphenated forms (components are separated by hyphen(s)), and satisfy semantic non-compositionality, statistical co-occurrence or literal phenomena [28] etc. Agarwal et al. (2004) classified the Bengali MWEs in three main classes using subclasses. Instead, we propose seven broad classes of Bengali compound noun MWEs considering their morpho-syntactic flexibilities, as follows:

- **Named-Entities (NE):** Names of people (*Rabindranath Thakur*, 'Rabindranath Tagore'), names of locations (*Bharat-barsa*, 'India'), names of organizations (it Pashchim Banga Siksha Samsad, 'West Bengal Board of Education') etc. where inflection is only allowed to be added to the last word.

- **Idiomatic Compound Nouns:** These are non-productive[2] 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 absence of inflection from the first component (*maa-baba*, 'mother and father').

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

- **Numbers:** They are highly productive, impenetrable and allow slight syntactic variations like inflections. Inflection 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 consist mostly of two tokens. Inflection can be added to the last word *pistuto bhai*, 'maternal cousin').

- **Conventionalized Phrases:** Sometimes, they are called as 'Institutionalized phrases'. Although not necessarily idiomatic, 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 that they are semi-productive (*hater panch*, 'remaining resource').

- **Reduplicated Terms:** Reduplications are non-productive and tagged as noun phrases. They are further classified as onomatopoeic expressions (*khat khat*, 'knocking'), complete reduplication (*bara-bara*, 'big big'), partial reduplication (*thakur-thukur*, 'God'), semantic reduplication (*matha-mundu*, 'head'), correlative reduplication (*maramari*, 'fighting') [11].

Identification of reduplication has already been carried out using the clues of Bengali morphological patterns [11]. A number of research activities in Bengali Named Entity (NE) detection have been conducted [23], but the lack of publicly available standard tools to detect NEs inhibits the incorporation of them within the existing system. Therefore, we discard the identification of NEs from this experiment. Kin terms and numbers can be easily captured by some well-developed lexicons because they are small in number and form a closed set in Bengali [2]. The present work mainly focuses on the extraction of productive and semi-productive bigrams, compound noun MWEs like idioms, idiomatic compound nouns, and simile terms (which are in open or hyphenated form) from a document using a semantic clustering technique.

3 Semi-automated approach for candidate extraction

3.1 Corpus acquisition and bigram extraction

Resource acquisition is one of the challenging obstacles to work with electronically resource constrained languages like Bengali. However, we crawled a large number of Bengali articles written by the noted Indian Nobel laureate Rabindranath Tagore[3]. While we are primarily interested in token level or phrase level characteristics, document information (e.g., the order of the documents, variation of the size of the documents, length normalization etc.) has not been maintained and manipulated in the experiment. Therefore, we merged all the articles and prepared a raw corpus consisting of 393,985 tokens and 283,533 types. The actual motivation for choosing the literature domain

---

[2]A phrase is said to be "productive" if new phrases can be formed from the combinations of syntactically and semantically similar component words of the original phrase.

[3]http://www.rabindra-rachanabali.nltr.org

in the present task was to obtain useful statistics to further help Stylometry analysis [9]. However in literature, the use of MWEs is greater than in the other domains like tourism, newspapers, scientific documents etc. because the semantic variability of MWEs offers writers more expressive terms. In Bengali literature, idiomatic expressions and relations terms are quite frequently used.

Since the preliminary crawled corpus was noisy and un-formatted, we used a basic semi-automatic pre-processing technique to make the corpus suitable for parsing. We used a Bengali shallow parser[4] to identify the POS, chunk, root, inflection and other morphological information of each to-ken. We observed that some of the tokens were misspelled due to typographic and phonetic errors. Thus, the Shallow parser could not be able to detect the actual root and in-flection of these two variations. To make the system fully automated, we allowed retaining the types of variations into the cleaned text.

After pre-processing, bigram noun sequences whose constituents were in the same chunk were extracted using their POS and chunk categories. We observed that dur-ing the parsing phase, the Shallow parser could not disam-biguate common nouns ('NN') and proper nouns ('NNP') appropriately. The reason could be the continuous need to coin new terms for new concepts. We took both of them and manually filtered the named-entities from the collected list so that we could accumulate most of the proper nouns for our main experimental module. Although the chunk infor-mation helps to identify the boundary of a phrase, some of the phrases belong to chunks having more than two nouns. The frequency of these phrases is also identified during the evaluation phase. Now, a bigram nominal candidate phrase can be thought of as $< M1\ M2 >$. The morphological heuristics used to separate the candidates are described in Table 1. After the first phase, a list of possible candidates was collected which was fed into the annotation phase.

Heuristics	
POS	POS of each bigram must be either 'NN' or 'NNP'
Chunk	M1 and M2 must be in the same 'NP' chunk
Inflection	Inflection[5] of $M1$ must be 'null', (-r), (-er), (-e), (-y) or (-yr)

Table 1: Heuristics for the candidate selection

3.2 Annotation study

Three anonymous annotators – linguistic experts working on our project – were hired to carry out the annotation. They were asked to divide all extracted phrases into four classes and definitions of the classes using the following definitions:
Class 1: Valid NN MWEs (M): phrases which show total non-compositionality and their meanings are hard to pre-dict from their constituents; e.g., *hater panch* ('remaining resource').
Class 2: Valid NN semantic collocations but not MWEs (S): phrases which exhibit partial or total compositionality

(e.g., act as institutionalized phrases) and show statistical idiomaticity; e.g., *bibaha barsiki* ('marriage anniversary').
Class 3: Invalid collocations (B): phrases enlisted due to bigrams in an n-gram chunk having more than two compo-nents; e.g., *porbot sohorer*, ('of mountain town').
Class 4: Invalid candidates (E): phrases enlisted due to the error in parsing like POS, chunk, inflection including named-entities; e.g., *granthagar tairi* ('build library').

Class 3 and class 4 types were filtered initially and their individual frequencies are noted as 24.37% and 29.53% re-spectively. Then the remaining 46.10% (628 phrases) of the total candidates were annotated and labeled as MWE (M) or S (Semantically collocated phrases), and they were fed into the evaluation phase. We plan to make the dataset publicly available soon.

The annotation agreement was measured using standard Cohen's kappa coefficient (κ) [16]. It is a statistical mea-sure of inter-annotation agreement for qualitative (categor-ical) items. It measures the agreement between two raters who separately classify items into some mutually exclu-sive categories. We employ another strategy in addition with kappa (κ) to calculate the agreement between anno-tators. We choose the measure of agreement on set-valued items ($MASI$) [29] that is used for measuring agreement in the semantic and pragmatic annotations. $MASI$ is a distance between sets whose value is 1 for identical sets, and 0 for disjoint sets. For sets A and B, it is defined as: $MASI = J * M$, where the Jaccard metric (J) is:

$$J = \frac{A \cap B}{A \cup B} \qquad (1)$$

Monotonicity (M) is defined as follows:

$$M = \begin{cases} 1, & if\ A = B \\ 2/3, & if\ A \subset B\ or\ B \subset A \\ 1/3, & if\ A \cap B \neq \phi, A - B \neq \phi\ \&\ B - A \neq \phi \\ 0, & if\ A \cap B = \phi \end{cases} \qquad (2)$$

The inter-annotation agreement scores of three annotators are presented in Table 2. Among the 628 types of noun-noun candidates, half of them selected randomly were used in the development phase and the remaining were used in the testing phase.

4 Baseline system

As mentioned earlier, the task of identifying Bengali com-pound nouns from a document has had little attention in the literature, and thus there is no prior developed methodol-ogy that can be used for the baseline. Therefore, in this ex-periment, we simply adapt a heuristic to develop our base-line system. The phrases which do not affix any nominal chunk and determinant at the prefix and suffix positions are selected as MWEs in the baseline system. The baseline system naturally reaches high accuracy in terms of recall since most of the identified MWEs satisfy the heuristics mentioned above. But in terms of precision, it shows very

MWEs [# 628]	Agreement between pair of annotators			
	A1-A2	A2-A3	A1-A3	Average
KAPPA	87.23	86.14	88.78	87.38
MASI	87.17	87.02	89.02	87.73

Table 2: Inter-annotation agreement

low accuracy (38.68%) since many collocated and fully-compositional elements were wrongly identified as MWEs. The main challenge of our model was to filter these irrelevant collocations from the selected candidate set.

5 Statistical methodologies

We started our experiment with the traditional methodology of collocation detection. Previous literature [15] [21] [26] shows that various statistical methodologies could be incorporated in identifying MWEs from a large corpus. In this experiment, we developed a statistical system using these previous techniques and modified them according to our requirements[6]. It is worth noting that frequency information of the candidate phrases in a corpus is a strong clue for labeling them as MWEs since it provides the evidence of more certainty of occurrence than randomness. However, for a resource-constrained language like Bengali, infrequent occurrence of candidates may not give any reasonable conclusion to judge them as MWEs (or Non-MWEs) because the size of the corpus itself is generally not adequate for statistical analysis. Therefore, instead of taking the frequency information directly, we took five standard association measures namely Point-wise Mutual Information (PMI) [15], Log-Likelihood ratio (LLR) [21], Co-occurrence measure [2], Phi-coefficient and Significance function [2] for extracting NN Multiword Expressions. A combined weighted measurement is proposed for the identification task, which is helpful to compute bigram collocation statistics. We ranked the list individually based on each of the statistical measures. We noticed in the comparative study that the results obtained by the frequency-based statistics like PMI and LLR could not identify MWEs at the top position of the ranked list. Therefore, we posited that the lexico-semantic affinity among the constituents could unleash the dependency of frequency information in the measurement. Final evaluation combined all the statistical features mentioned above. Experimental results on the development dataset show that Phi-coefficient, Co-occurrence and Significance functions which are actually based on the principle of collocation produce more accurate results compared to direct frequency-based measurements like LLR, PMI in the higher ranks. So, these three measures are considered in the weighted scheme to assign certain weights to the candidate phrases. After a continuous weight tuning over the development data, the best weights for Co-occurrence, Phi and Significance functions are re-
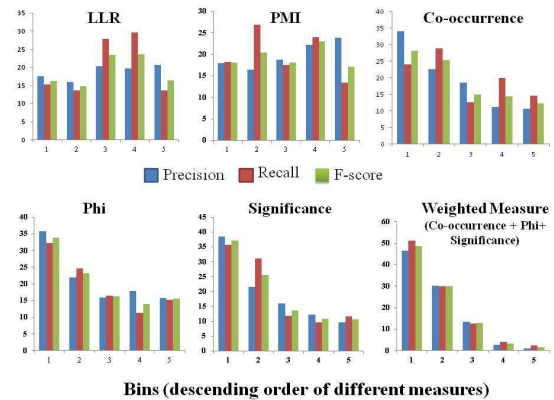

Figure 1: Performance of all the statistical measures and the weighted measure.

ported as 0.45, 0.35 and 0.20 respectively for the combined measurement. The individual score of each measure is normalized before assigning weights so that they fall in the range of 0 to 1. For each measurement, the scores have been sorted in descending order and the total range is divided into five bins (bin 1 signifies higher ranked bin). The intuition is that the more the value of the statistical measure for a candidate phrase, the more it behaves like a MWE. The metrics used to evaluate the statistical systems are described below:

Precision in bin i (P_i) = (Number of MWEs present in the ith bins) / (total number of candidates in ith bins)

Recall in bin i (R_i) = (Number of MWEs present in the ith bins) (total number of MWEs in the documents)

F-score in bin i $(F_i$) = $(2 * P_i * R_i)(P_i + R_i)$

Figure 1 shows the results obtained from five association measures and the combined weighted measures over the test dataset.

6 Semantic clustering approach

Multiword Expressions represent a core semantic issue that can be partially resolved by morphological or statistical clues. However, it often fails at capturing the underlying semantic notion of forming a new multiword expression, i.e., the meaning of the entire expression cannot be predicted by aggregating the meaning of its components. Our proposed approach aims to handle these drawbacks by considering individual senses induced by the components of an expression. This approach tries to cluster semantically related words present in the document. However,

---

[6]Interested readers are encouraged to go through the research dissertation by [10]

for a particular token present in the document, finding semantically similar words appeared in the corpus can be carried out by looking at the surroundings tokens and finding the synonymous entries of the surrounding words within a fixed context window. However in that case, a high number of occurrences of a particular token should be needed in a corpus in order to obtain statistically significant evidences. Therefore, in a medium-size corpus, it is hard to extract the cluster of synonyms. Since the electronic resources such as newspapers, weblogs may not be present for all the languages and the presence of frequent MWEs in such contents are rare, we focus on extracting the MWEs only from the medium size crawled corpus. However, semantics of a word may be obtained by analyzing its similarity set called the synset that indeed expresses its meaning in different contexts. Therefore, semantic distance of two tokens in a phrase can be measured by comparing their synsets properly. Higher value of the similarity between two sets indicates semantic closeness of two tokens to each other. For instance, let $M1$ and $M2$ be two components of a bigram $< M1\ M2 >$. For each component of the expression, semantically related words present in the documents are extracted by using the formatted Bengali monolingual dictionary (discussed in Section 6.1) and two separate clusters are formed for two tokens. Now, intersection of two clusters can be a suitable measure to judge the commonality of two components appeared in a bigram. Using these common elements, three different similarity measurements are proposed in our algorithm and the results are reported separately in Table 5 later. Finally, based on a predefined threshold, the candidate phrases were labeled as MWE or Non-MWE.

6.1 Restructuring the Bengali monolingual dictionary

To the best of our knowledge, no full-fledged WordNet or thesaurus is available in Bengali. In this section, we describe the construction of a Bengali thesaurus that aims not only to develop Bengali WordNet but also to identify the meaning of multiword expressions. Focusing mainly on MWEs, the present natural language resource is being developed from the available Bengali-to-Bengali monolingual dictionary (Samsada Bengali Abhidhana[7]). The monolingual dictionary contains each word with its parts-of-speech (Noun, Adjective, Pronoun, Indeclinable, Verb), phonetics and synonym sets. Synonym sets are separated using distinguishable notations based on similar or differential meaning. Synonyms of different sense with respect to a word entry are distinguished by a semicolon (;), and synonyms having same sense are separated by a comma (,). An automatic technique is devised to identify the synsets for a particular word entry based on the clues (, and ;) of similar and differential senses. The symbol tilde (~) indicates that the suffix string followed by the tilde (~) notation

---

[7] http://dsal.uchicago.edu/dictionaries/biswas-bangala/

makes another new word concatenating with the original entry word. A snapshot of the modified synset entries of the Bengali word *Angshu* is shown in Figure 2. Table 3 shows the frequencies of different synsets according to their part-of-speech.

Figure 2: Monolingual dictionary entry and built synsets for the word *Angshu*.

6.2 Generating synsets of nouns

At the beginning of the clustering method, we generate a synonym set for each noun present in the corpus using the modified dictionary. However, the formatted dictionary can be assumed to be a close set of word entries $W^1, W^2, W^3, ..., W^m$ where the synsets of the entries look like:

$$W^1 = n_1^1, n_2^1, n_3^1, ... = n^1$$
$$W^2 = n_1^2, n_2^2, n_3^2, ... = n^2$$
$$...$$
$$W^m = n_1^m, n_2^m, n_3^m, ... = n^m$$

where $W^1, W^2, ..., W^m$ are the dictionary entries and n^i denotes the set of synsets of the entry W^i. Now each noun entry identified by the shallow parser in the document is searched in the synset entries of the dictionary for its individual existence with or without inflection. For instance, N is a noun in the corpus and it is present in the synsets of W^1, W^3 and W^5. Therefore, they become entries of the synset of N. Formally, this can be represented as follows.

$$Synset(N) = \{W^l, W^3, W^5\} \qquad (3)$$

Equation 2 states that since the given noun N is present in the synsets of W^1, W^3 and W^5, the sense of these three dictionary entries are somehow related to the sense of N. Following this, the synonym noun tokens for each of the nouns present in the corpus are extracted from the dictionary. In short, the formatted dictionary indeed helps us cluster synonymous tokens corresponding to a particular noun present in a document.

Word Entries	Synset	Noun	Adjective	Pronoun	Indeclinable	Verb
47949	63403	28485	11023	235	497	1709

Table 3: Frequency information of the synsets with different part-of-speeches.

6.3 Semantic relatedness between noun synsets

The next task is to identify the similarity between the synsets of two nouns that can help measure the semantic relatedness between them. This is done by taking the intersection of the synsets and assigning a score to each such noun-pair to indicate the semantic affinity between two nouns. For instance, if N_i and N_j are two nouns in the document, and S_i and S_j are their corresponding synsets extracted using the technique stated in Section 6.2, then the commonality of the two nouns can be defined as:

$$Comm(N_i, N_j) = |S_i \cap S_j| \qquad (4)$$

The above equation shows that the commonality is maximum when the similarity is measured with itself (i.e., $Comm(N_i, N_j)$ is maximum when $i = j$).

6.4 Semantic clustering of nouns

Using the scores obtained by the semantic commonality measure discussed in the previous subsection, we can build a cluster centered on a given noun present in the document such that the cluster constitutes all the nouns semantically related to the given noun (discussed in subsections 6.2 and 6.3). A score is assigned to each such noun present in the cluster representing the semantic similarity (discussed in subsection 6.3) between this noun and the noun present at the center of the cluster. An illustrative example is shown in Figure 3. For example, suppose the nouns identified by the Shallow parser in the document are
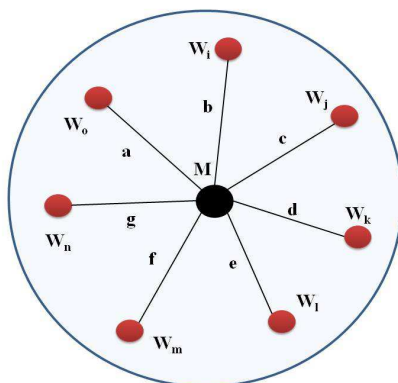

Figure 3: Semantic clustering for a given noun M and the associated commonality scores with similar nouns (W_i, W_j,..., W_o etc). In this figure, the semantic similarities of M with the other nouns are denoted by the weights (i.e., a, b, c etc.) of the edges.

6.5 Decision algorithm for identifying MWEs

We extract the candidates eligible for judging MWE in section 3. The elaborated algorithm to identify a noun-noun bigram (say, $< M1\ M2 >$) as MWEs is discussed below with an illustrative example shown in Figure 3.

Algorithm: MWE_CHECKING
Input: Noun-noun bigram $< M_1\ M_2 >$
Output: Return true if MWE, or return false.
1. Extract semantic clusters of M_1 and M_2 (discussed in Section 6.4);
2. Intersect the clusters of M_1 and M_2 (Figure 4 (left) shows the common synset entries (broken rectangles) of M_1 and M_2);
3. For measuring the semantic similarity between M_1 and M_2:
    3.1. In an n-dimensional vector space (n denotes the number of elements common in both the synsets of M_1 and M_2, e.g., in the Figure 4 (left), n=2), the common entries act as the axes. Put M_1 and M_2 as two vectors and their associated similarity with the axes tokens as their co-ordinates.
    3.2. Calculate cosine-similarity measurement and Euclidean distance between the two vectors (Figure 4 (right)).
4. Final decision is taken separately for two different measurements:
    4.1 If (cosine-similarity $> \alpha$) return true; else return false;
    4.2 If (Euclidean distance $> \beta$) return true; else return false;
    (where α and β are the pre-defined cut-off values determined from the development set)

Here, we elaborate step 3 and step 4 since the central theme of the algorithm lies in these two steps. After identifying the common terms from the synsets of the components of a candidate, a vector space model is used to identify the similarity between the two components. In n-dimensional vector space, these common elements denote the axes and each candidate acts as a point in the n-dimensional space. The coordinate position of the point (each component of the candidate bigram) in each direction is represented by the similarity measure between the synsets of each component and the noun representing the axis in that direction. The cut-off value for the classification of a given candidate as MWE (or Non-MWE) is determined from the development dataset after several tries to get the best performance (described in step 4). We have seen significant results for the cut-off values (0.4-0.6) on the development set based on F-sore measure. Therefore, we report the results on the test dataset for each of these cut-off values separately in Table 4. In the experiment, we observe that the bigrams that are actual MWEs, mainly non-compositional phrases, show a low similarity score between the synsets of their components.

If we take an example of the Bengali idiom – *hater panch* ('remaining resource'), we can see that English WordNet defines two components of the idiom in the following way: *hat* ('hand') as 'a part of a limb that is farthest from the torso' and *panch* ('five') as 'a number which is one more than four'. So from these two glosses it is quite evident
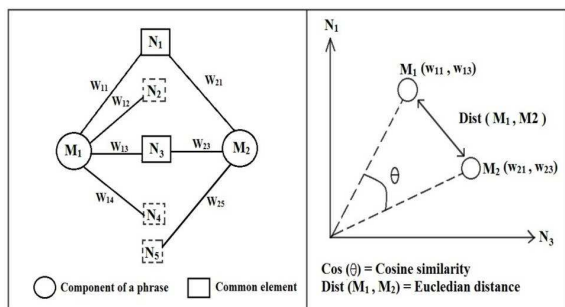
Figure 4: Intersection between the clusters of the components of a candidate bigram (left) and the similarity between two components (right)

that they are not at all semantically related. The synonym sets for these two components extracted from the formatted monolingual dictionary are as follows.

Synset (হাত) = { হস্ত, কর, পাণি, বাহু, ভুজ, কৌশল, হস্তক্ষেপ, ধারণ, রেখা, লিখিত, হস্তাক্ষর, হস্তান্তর, হাজা }
Synset (পাঁচ) = { পঞ্চ, সংখ্যা, কর্ম, গঙ্গা, গবা, কন্যা, গুণ, গৌড়, ভঙ্গ, তীর্থ, পক্ষদ্ব, পনেরো, পূর্ণিমা, পঞ্চাশ }

We can observe that the two synonym sets have no element in common and therefore their similarity score would be zero. In this case, a vector space model cannot be drawn in zero dimensional space. For them, a final concession weight is assigned to treat them as fully non-compositional phrases. To identify their non-compositionality, we need to show that their occurrences are not by mistake (i.e., because of a typo or due to unawareness of the author); rather they can occur side by side in several instances. But the concrete statistical proof can only be obtained using a large corpus. Here, for the candidate phrases which have zero similarity, we observe their existence more than one time in the corpus and then treat them as MWEs.

7 WordNet similarity measurement

We also incorporate English WordNet 2.1[8] in this experiment to measure the semantic distance between two Bengali words after translating them into English. Though the idea is trivial considering the manual intervention of the translation process, our main focus was to get an idea of how the semantic similarity of two components can help identify the combination as an MWE, and how a well-defined lexical tool is essential in the presently adapted linguistic environment. As already mentioned, WordNet::Similarity is an open-source package developed at the University of Minnesota for calculating the lexical similarity between word (or sense). Basically, it provides six measures of similarity and three measures of relatedness based on the WordNet lexical database [25]. The measures are based on the analysis of the WordNet hierarchy.

The measures of similarity are divided into two groups: path-based and information content-based. We chose two similarity measures in WordNet::Similarity for our experiments: WUP and LCH; WUP finds the path length to the root node from the least common subsumer (LCS) of the two word senses that is the most specific word sense they share as an ancestor [34]. In this experiment, we first translate the root of two Bengali components in a candidate phrase into their English forms using the Bengali-to-English Bilingual Dictionary[9]. Then these two words are run through the WordNet based Similarity module for measuring their semantic distance. A predefined cut-off value (μ) is determined from the development set to distinguish between an MWE and a simple compositional term. If the measured distance is less than the threshold, the similarity between them is less. The results are noted for different cut-off values as shown in Table 5. The bold font in each column shows the highest accuracy among different cut-off values.

8 Experimental results

We used standard IR metrics, i.e., Precision, Recall and F-score to evaluate the final results obtained from three similarity measuring modules (i.e., cosine-similarity, Euclidean distance and WordNet similairty) as discussed in the previous section. The evaluation of the systems was carried out on the previously mentioned hand-annotated dataset and the final results are shown in Table 5. The predefined threshold acquired from the development set was tuned to obtain the best results for all the similarity measures. Increasing recall with the increase of cut-off values indicates that most of the MWEs are identified across the wider threshold range. But the precision is not increasing gradually with the threshold. This result signifies that besides capturing most of the significant MWEs, it also considers more false positives at higher cut-off values. Our goal is to pick up an optimal point where both precision and recall stabilize with the reasonable results and minimize the erroneous predictions. The Cosine-similarity [?] achieves maximum precision at 0.5, whereas Euclidean distance and WordNet::Similarity achieve maximum precision at 0.4 and 0.5 respectively. The effect of English WordNet in identifying Bengali MWEs is noticeable in Table 5. Wordnet::Similarity identifies the maximum number of MWEs correctly at the cut-off of 0.5. Baldwin et al. (2003) suggested that WordNet::Similarity measure can be used to identify Multiword Expression decomposability. This is once again effective for Bengali MWE identification. There are also candidates with very low value of similarity between their constituents (e.g., *ganer jagat* (earth of song, 'affectionate of song')), yet they are discarded from

---

[8]http://www.d.umn.edu/~tpederse/similarity.html

[9]http://dsal.uchicago.edu/dictionaries/biswas-bengali/

Cut-off	Cosine-Similarity			Euclidean Distance			WordNet::Similarity		
	P	R	F	P	R	F	P	R	F
0.6	70.75	64.87	67.68	70.57	62.23	66.14	74.60	61.78	67.58
0.5	78.56	59.45	67.74	72.97	58.79	65.12	80.90	58.75	68.06
0.4	73.23	56.97	64.08	79.78	53.03	63.71	75.09	52.27	61.63

Table 4: Precision, Recall and F-score of three measures (in %) in clustering approach and WordNet::similarity measure.

this experiment because of their low frequency of occurrence in the corpus that could not reveal enough evidence of considering them as MWEs. Whether such an unexpectedly low frequent high-decomposable elements warrant an entry in the lexicon depends on the type of lexicon being built [4].

9 Discussion

At the beginning of the article, we claimed that the increasing degree of semantic similarity between two constituents of a candidate bigram indicates the less probability of the candidate to be a multiword expression. The statistical methodologies achieve low accuracy because the medium size corpus fails to unfold significant clue of their occurrences to label the non-compositional phrase as MWEs. We have adopted an approach taking into account the semantic interpretation of MWE that seems to be unconventional in the task of identifying MWEs in any language. In the experimental results, the semantic clustering approach outperforms the other systems. However, the clustering algorithm is able to identify those MWEs whose semantics are fully opaque from the semantics of their constituents (strictly non-compositional). But MWEs show a continuum spectrum from fully-compositional (e.g., idioms) to institutionalized phrases (e.g., traffic signal) where high statistical occurrence is the only clue to identify them as MWEs. These partial or transparent expressions are not captured by our system because of the lack of a large size standard corpus. The presence of the monolingual dictionary is another important criterion to carry out the proposed approach. It acts as a proxy for an individual noun to cumulate the related noun tokens. This algorithm assumes that every language should possess its own dictionary since it is the first and fundamental resource used not only for experimental purposes but also for language generation and understanding.

10 Conclusion

We hypothesized that sense induction using synonym set can assist in identifying multiword expressions in Bengali. We introduced a semi-automated approach to establish the hypothesis. We compared our results with the baseline system and the traditional statistical systems. We have shown that clustering measure can be an effective measure to enhance the extraction task of MWEs. The contributions of

the paper are fourfold: firstly, we provide an efficient way of clustering noun tokens having similar sense; secondly, we propose a semantic similarity based approach for identifying MWEs; thirdly, it a preliminary attempt to reconstruct a Bengali monolingual dictionary as a standard lexical thesaurus and finally, the present task is a pioneering step towards the development of Bengali WordNet. At last, we would like to stress that this entire methodology can be used to identify MWEs in any other language domain. In the future, we plan to extend the algorithm to support all ranges of compositionality of Bengali MWEs. Moreover, we modify the semantic interpretation of MWEs to enlist partial and compositional phrases as much as possible. Furthermore, incorporating the Named-Entity recognizer can help develop a full-fledged MWE identification system. Finally, we will make the formatted monolingual dictionary publicly available soon and incorporate the strictly non-compositional MWEs which rarely occur in the medium-size corpus into the dictionary so that they are directly captured from the thesaurus.

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] Agarwal, A., Ray, B., Choudhury, M., Sarkar, S., Basu, A. (2004). Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. *In Proc. of International Conference on Natural Language Processing (ICON)*, pp. 165-174.

[3] Agirree, E., Aldezabal, I., Pociello, E. (2006). Lexicalization and multiword expressions in the Basque WordNet. *In Proc. of Third International WordNet Conference*. Jeju Island (Korea).

[4] Baldwin, T., Bannard, C., Tanaka, T., Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. *In Proc. of the Association for Computational Linguistics, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 89–96.

[5] Baldwin, T., Kim, S. N. (2010). Multiword Expressions, in Nitin Indurkhya and Fred J. Damerau

(eds.) *Handbook of Natural Language Processing*, Second Edition, CRC Press, Boca Raton, USA, pp. 267—292.

[6] Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1). pp. 23-35.

[7] Bhtnariu, C., Kim, S.N., Nakov, P., Oseaghdha, D., Szpakowicz., S., Veale, T. (2009). SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. *In Proc. of the NAACL Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009) at NAACL*, pp. 100-105.

[8] Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C., Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. *In proc. of the Third International Conference on Language Resources and Evaluation (LREC)*, pp. 1934–1940.

[9] Chakraborty, T. (2012). Authorship Identification in Bengali Literature: a Comparative Analysis. *InProc. of 24th International Conference on Computational Linguistics (Coling, 2012)*, Mumbai, India, pp. 41-50.

[10] Chakraborty, T. (2012). *Multiword Expressions: Towards Identification to Applications*, LAP LAMBERT Academic Publishing GmbH & Co., KG Heinrich-Bocking-Str. 6-8, 66121, Saarbrucken, Germany, ISBN 978-3-659-24956-3.

[11] Chakraborty, T., Bandyopadhyay, S. (2010). Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach. *In proc. of the 23rd International Conference on Computational Linguistics (COLING 2010), Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*. Beijing, China, pp 72-75.

[12] Chakraborty, T., Das, D., Bandyopadhyay, S. (2011). Semantic Clustering: an Attempt to Extract Multiword Expressions in Bengali. *In Proc. of Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, Portland, Oregon, USA, pp. 8-11.

[13] Chakraborty, T., Pal, S., Mondal, T., Saikh, T., Bandyopadhyay, S. (2011). Shared task system description: Measuring the Compositionality of Bigrams using Statistical Methodologies. *In Proc. of Distributional Semantics and Compositionally (DiSCo), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, Portland, Oregon, USA, pp. 38 – 42.

[14] Chattopadhyay, S. K. (1992). Bhasa-Prakash Bangala Vyakaran, Third Edition.

[15] Church, K. W., Hans, P. (1990). Word Association Norms, Mutual Information and Lexicography. *In Proc. of 27th Association for Computational Linguistics (ACL)*, vol. 16(1), pp. 22-29.

[16] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol. 2, pp. 37–46.

[17] Das, D., Pal, S., Mondal, T., Chakraborty, T., Bandhopadhyay, S. (2010). Automatic Extraction of Complex Predicates in Bengali. *In Proc. of Multiword Expressions: from Theory to Applications (MWE 2010), The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, pp. 37-45.

[18] Dasgupta, S., Khan, N., Sarkar, A. I., Pavel, D. S. H., Khan, M. (2005). Morphological Analysis of Inflecting Compound Words in Bengali. *In Proc. of the 8th International Conference on Computer and Information Technology (ICCIT)*, Bangladesh.

[19] Diab, M., Bhutada, P. (2009). Verb noun construction MWE token supervised classification. *In proc. of the Workshop on Multiword Expressions*, Singapore, pp. 17-22.

[20] Dias, G. H. (2003). Multiword Unit Hybrid Extraction. *In proc. of the Association for Computational Linguistics, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 41-48.

[21] Dunning, T. (1993). Accurate Method for the Statistic of Surprise and Coincidence. *In Computational Linguistics*, pp. 61-74.

[22] Dwork, C., Kumar, R., Naor, M., Sivakumar, D. (2001). Rank aggregation methods for the web. *In proc. of Conference on the World Wide Web (WWW)-ACM*, New York, pp. 613-622.

[23] Ekbal, A., Haque, R., Bandyopadhyay, S. (2008). Maximum Entropy Based Bengali Part of Speech Tagging. *In proc. of Advances in Natural Language Processing and Applications Research in Computing Science*, pp. 67-78.

[24] Fazly, A., Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. *In proc. of Association for Computational Linguistics, Workshop on a Broader Perspective on Multiword Expressions*. Prague, Czech Republic, pp. 9-16.

[25] Fellbaum, C. (1998). *WordNet: An Electronic lexical Database*. MIT Press, Cambridge, USA.

[26] Kilgarriff, A., and Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities. Senseval Special Issue*, 34(1-2), pp. 15-48.

[27] Korkontzelos, I, Manandhar, S. (2009). Detecting Compositionality in Multi-Word Expressions. *In proc. of the Association for Computational Linguistics-IJCNLP*, Singapore, pp. 65-68.

[28] Kunchukuttan, F. A., Damani, O. P. (2008). A System for Compound Noun Multiword Expression Extraction for Hindi. *In proc. of 6th International Conference on Natural Language Processing (ICON)*. pp. 20-29.

[29] Passonneau, R. J. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. *Language Resources and Evaluation*.

[30] Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *In Proc. of Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pp. 1-15.

[31] Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24 (4), pp. 35–43.

[32] Tanaka, T., Baldwin, T. (2003). Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing. *In Proc. of the Association for Computational Linguistics, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 17–24.

[33] Venkatpathy, S., Joshi, A. (2009). Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. *In Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Association for Computational Linguistics*, pp. 899–906.

[34] Wu, Z., Palmar, M. (1994). Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133–138.