Design and Development of a Bangla Semantic Lexicon and Semantic Similarity Measure

Manjira Sinha
Department of Computer Science and Engineering
Indian Institute of Technology

Tirthankar Dasgupta
Department of Computer Science and Engineering
Indian Institute of Technology

Abhik Jana
Department of Computer Science and Engineering
Indian Institute of Technology

Anupam Basu
Department of Computer Science and Engineering
Indian Institute of Technology

ABSTRACT
In this paper, we have proposed a hierarchically organized semantic lexicon in Bangla and also a graph based edge-weighting approach to measure semantic similarity between two Bangla words. We have also developed a graphical user interface to represent the lexical organization. Our proposed lexical structure contains only relations based on semantic association. We have included the frequency of each word over five Bangla corpuses in our lexical structure and also associated more details to words such as, whether the words are mythological or not, whether it can be used as verb or not, in order to use the word as a verb which word should be appended to it etc. As we have earlier discussed, this lexicon can be used in various applications like categorization, semantic web, and natural language processing applications like, document clustering, word sense disambiguation, machine translation, information retrieval, text comprehension and question-answering systems.

General Terms
Semantic lexicon development

Keywords
Bangla SynNet , Semantic Similarity, Category, Concept, Sub-concept, Cluster

1. INTRODUCTION
The *lexicon* of a language is a collection of lexical entries consisting of information regarding words and expressions. According to Levelt [8] every lexical entry contains mainly two types of information namely, *form* and *meaning* that help a user to recognize and understand words. *Form* refers to the orthography, phonology and morphology of the lexical item and *Meaning* refers to its syntactic and semantic information.

A lexicon is the central part of any natural language processing applications like, machine translation, language comprehension, language generation and information retrieval. Depending on the storage structure and the content, the type of lexicon varies. For example, dictionary, thesaurus, FrameNet, WordNet and ConceptNet are different type of lexicons having different lexical representation schemes. One of the most popular and commonly used lexicons in the present time is the WordNet that organizes words in terms of their senses. Given the importance and wide acceptability of WordNet in computational linguistics, several attempts have been made to develop such lexical representation schemes for many other languages[1]. Attempts have also been made to develop WordNet like lexical representation scheme in Indian languages. One of the widely known such work is the Hindi WordNet [30]. However, developing such a complex lexical representation scheme is not trivial. It not only requires an extensive linguistic expertise, but also manual encoding of the individual synsets need a huge time and manual effort. As a result of this, a lot of attempts are currently going on to develop semi-supervised algorithms to compute semantic distance between words.

Bangla is an Indo-Aryan language. It is native to the region of eastern South Asia known as Bengal, which comprises present day Bangladesh, the Indian state of West Bengal, and parts of the Indian states of Tripura and Assam. It is written using the Bengali script. It has been estimated that about 230 million people in the world speaks Bangla which is the sixth most spoken language in the world[2]. Bangla is also the national language of Bangladesh. Despite being so popular, very few attempts have been made to build a semantically organized lexicon of substantial size in Bangla [16]. But, from the above discussions, it is evident that a semantic lexical representation is essential to the development of number of NLP applications in Bangla.

In this paper we present the design and development of a Bangla lexicon that is based on the semantic similarity among Bangla words. The lexicon can be further used in various applications like as mentioned above. The design of this structure is based on *Samsad Samarthak Sabdokosh* [11]. The lexicon is based on a hierarchical organization where at the top there is a root node which is divided into different *categories*. The *categories* are divided into *concepts*. The *concepts* are divided *sub-concepts* which are further divided into *clusters*. The words are grouped into *clusters* along with their synonyms. Each *category*, *concepts*, *sub-concepts* and *clusters* are connected in terms of weighted edges. The weight denotes the semantic distance between the two nodes connected by an edge. All together the lexicon contains more than 50,000 unique Bangla words connected in terms of their semantic similarities.

1 http://globalwordnet.org/

2 http://en.wikipedia.org/wiki/Bengali_language

Based on the hierarchical representation of our lexicon, we have developed a semantic similarity measure between Bangla words. The similarity measure was evaluated by a number of native Bangla speakers where we have achieved a significantly high accuracy.

The rest of the paper is as organized follows: Section 2 contains background study. We have also pointed out some of the differences of our proposed structure with respect to WordNet in Section 2. Section 3 explains the design and implementation of the lexicon along with some details of basis of our lexicon. Section 4 describes the proposed approach of predicting semantic similarity between words; in Section 5, we have discussed about the evaluation of our proposed semantic similarity method. Finally, we conclude this paper in section 6.

2. BACKGROUND STUDY

Plethora of works has been done developing semantic lexicons in various languages like, English, French, Dutch, German and Italian3. The efforts ranges from developing lexicons like, Dictionary, and Thesaurus, to more advance forms like, WordNet, CYC and others. Synsets are main building block of such lexical representations. A list of synonymous word forms a synset which are further connected in terms of different semantic relations like, Hyponym, Hypernym, Holonym, Metonym, and Meronym. These relations are nothing but semantic pointers.

With respect to this, surprisingly, very few attempts have been taken to develop semantic lexicons in language like Bangla which is among top ten most spoken languages in the world. The Bangla WordNet project [4] is the only such attempt that aims to build a large scale semantic lexicon for Bangla words. However, at the present state the lexicon is reported to compose of around 36000 words as compared to our proposed lexicon of 50,000 words. Further, the structure is based on Bangla to English bi-lingual dictionaries and in strict alignment (only the synonym equivalents are used) with the Princeton WordNet for English.

Our proposed lexical representation SynNet is different from WordNet in many aspects. Some of the important differences being, SynNet contains cross part-of speech links which are not present in WordNet; it contains semantic relations such as "actor"([book]-[writer]), "instrument"([knife]-[cut]); the links are weighted to indicate the measure of semantic similarity between any two pair of words and moreover, SynNet acts as an thesaurus for Bangla rather than like a Dictionary.

2.1 Works on measuring semantic similarity among words

A number of approaches for measuring conceptual similarity have been taken in the past. Tversky's feature based similarity model [20], is among the early works in this field. Some works [13,6,7] have proposed the conceptual distance approach that uses edge weights, between adjacent nodes in a graph as an estimator of semantic similarity. Resnick [14, 15] have proposed the information theoretic approach to measure semantic similarity between two words. Here, the class is made up of all words in a noun synset as well as words in all directly or indirectly subordinate synsets. Conceptual similarity between two classes is approximated by the information content of the first class in the noun hierarchy that subsumes both classes. Richardson et al. [16] has proposed an edge-weight based scheme for Hierarchical Conceptual Graphs (HCG) to measure semantic similarity between words. According to them, the weight of a link depends on three factors: the density of the HCG at that point, depth in the HCG and the strength of connotation between parent and child nodes. Efforts (JC 1997) have been made to combine both the information content based approach and the graph based approach of predicting semantic similarity. In addition, strategies of using multiple information sources to collect semantic information have also been adopted [9]. Wang [21] have criticized the traditional notions of the depth and density in a lexical taxonomy. They have proposed novel definitions of the depth and density which have found to give significant improvement in performance; they have also verified the results with human judgements. However, almost all of the attempts described above have been taken in English based on the representation of WordNet. Das and Bandopadhaya [3] have proposed a Semantic Net in Bangla, where the relations are based on human pragmatics.

3. OVERVIEW OF THE PROPOSED LEXICON

We have taken the *Samsad Samarthak Sabdokosh* by Ashok Mukhopadhyay [4] as the basis for our proposed lexical representation in Bangla. In order to build up a semantic relation based lexical representation, we have constructed a hierarchical conceptual graph. An illustration of such a hierarchical representation scheme is depicted in Figure 1.

3http://globalwordnet.org/

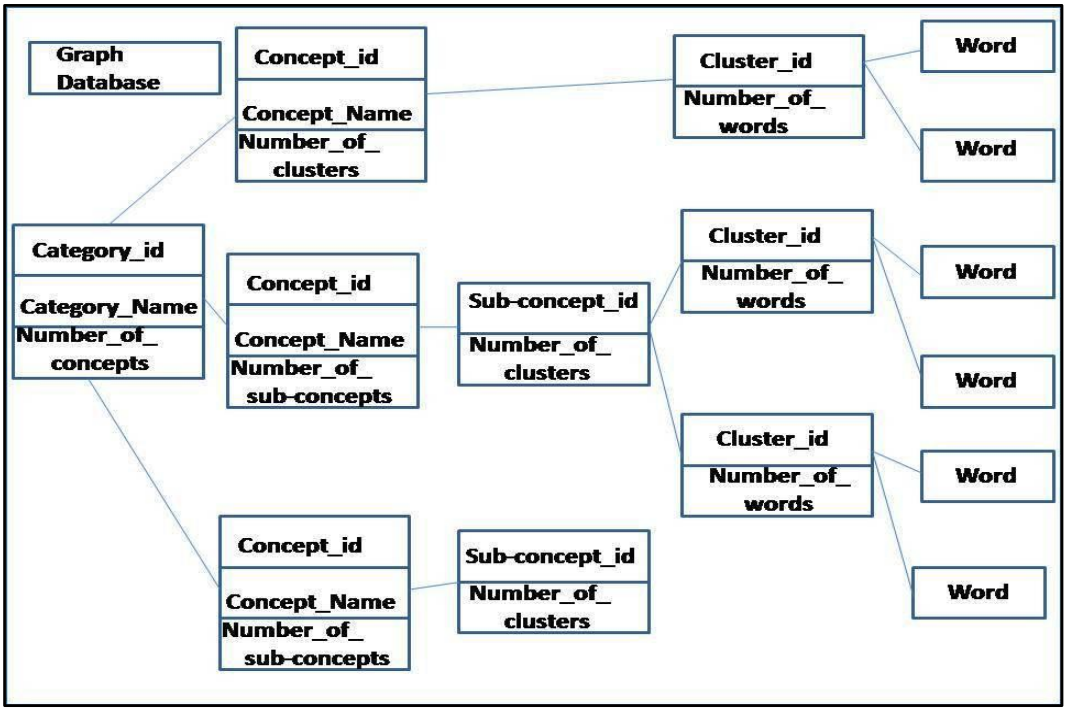

Figure 1: Schema of the proposed lexicon

We have classified each level of hierarchy in terms of "*domains*", "*concepts*", "*sub-concepts*", and "*clusters*". Accordingly, we have 30 different domains. Each domain is consisting of different concepts. The concepts are classified into sub-concepts. Different groups of words that are semantically related to a single sub-concept are organized together. Relevant information such as Part-Of-Speech (POS) corresponding to every word and antonyms for adjectives are also mentioned. Concepts are further classified into clusters. Each cluster is consisting of semantically similar words which are further grouped according to their degree of semantic similarity thus, making the whole structure hierarchical in nature. We have used different markers to separate out each cluster as well as words within each cluster. Each word in our lexicon is composed of the following **11-tupules**:

1 Word
2 Corpus frequency: computed from a Bangla lexicon of 35 million words.
3 Cluster_id: Id of the cluster of which the word is a member
4 Part-Of-Speech (POS)
5 Concept_id: Id of the concept in which the word belongs
6 Sub-concept_id: Id of the sub-concept(if exists) in which the word belongs
7 Category_id: Id of the category in which the word resides
8 Myth: a flag to indicate any mythical relation to the word
9 Antonym: a flag to indicate whether the word is antonym of the concept or not
10 Is_collective: a flag to indicate whether the word is a collective noun or not
11 G_word: a pointer to the general word denoting the collection in which the present word belongs
12 Verb: a flag to indicate whether the word can be also used as a verb or not.
13 To_verb: contains a word which can be appended to the present word to make it possible to be used as a verb. The no of word can be more than one also.
14 Primary link
15 Secondary link

In order to compute frequencies of each lexical item, we have prepared a Bangla corpuses composed of complete novel and story collection of Rabindranath Tagore, Bankimchandra Chattaopadhayay[4], collection of Bangla blogs over the internet, Bangla corpus by CIIL Mysore[5] and Anandabazar news corpus[6]. All together there are 35 million words from which we have prepared a list of around 4 lakh distinct words in Bangla with their corpus frequencies.

Given a word, its frequency over the five mentioned corpuses, its associations with different categories or sub-categories are collected at a single place so that a user can navigate through the storages with low cognitive load. We have also rated the various types of connections among different levels of the graph and developed a mechanism for predicting semantic similarity measures between words in the proposed lexicon. It supports queries like DETAILS(X) (here X can be any type of node of the hierarchy) and SIMILARITY (WORD1, WORD2).

4 Both the Rabindra Rachanabali and Bankim Rachanabali documents is collected from www.nltr.org

5 Downloaded from www.ciil.org

6 Downloaded from www.cel.iitkgp.ernet.in

3.1 The Primary and Secondary Links

There are two types of cross links exists in our semantic lexicon - *primary links*, and *secondary links* which are the specified after words under clusters. The *primary link* refers to concepts or sub-concepts which are semantically very close to the word after which the link is specified for example the word গ্রহজগৎ/planetary system which is under মহাবিশ্ব/universe concept, has a *primary link* to the concept সূর্য/sun . The *secondary link* refers to concepts or sub-concepts which are somehow or in some generalized senses semantically related to the word after which the link is specified for example the word জ্যোতির্বিদ্যা/astrology which is under concept গ্রহ/planet has a secondary link to the concept নক্ষত্র/star. Primary link is represented by special tags like, <\primary>. After <\primary> a concept number or a sub-concept number is given to which the word has primary link. Secondary link is represented by the tag <\secondary>. In this case within the tags a concept number or sub-concept number is given to which the word has secondary link. The number of primary link or secondary link can be more than one also for a particular word. An illustration of the primary and secondary links in our lexicon is shown in Figure 2. In figure 2 below, the category id of মহাবিশ্ব–প্রকৃতি–পৃথিবী–গাছপালা/ universe-nature-earth-flora is 1, মহাবিশ্ব/ universe has sub-category id 1.1 meaning it is the 1st sub-category of category 1 and নিখিলভুবন/ universe cluster id 1.1.1 as it belongs to the synonym cluster of 1.1. The member relations of words with their clusters have been shown in dashed lines and the round dotted line and the compound line indicate primary link and secondary link respectively.

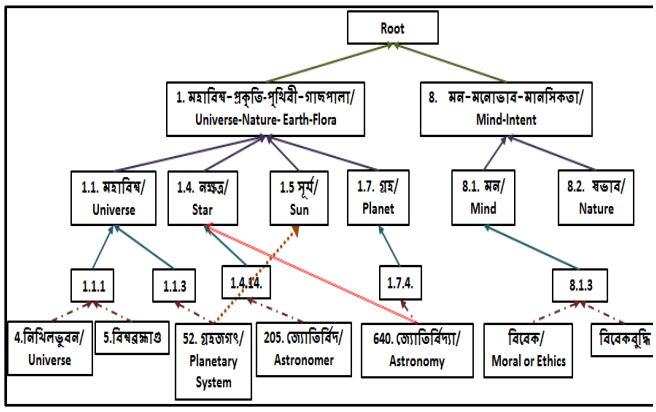

Figure 2: Hierarchical Representation of the Bangla Semantic Lexicon

3.2 Development of the Proposed Lexicon

In order to build-up a semantic relation based lexical representation Bangla; we have constructed a hierarchical conceptual graph based on the above mentioned thesaurus. We have also individually processed and stored the distinct general words in the book along with their respective details. Our storage and organization of the database facilitate computational processing of the information and efficient searching to retrieve the details associated with any word. Therefore, it will be a useful resource and tool to other psycholinguistic and NLP studies in Bangla. Given a word, its associations with different categories or sub-categories are collected at a single place so that a user can navigate through the storages with ease. We have also rated the various types of connections among different levels of the graph and developed a mechanism for predicting semantic similarity measures between words in the proposed lexicon. The details of the organizational methodology are described below.

As discussed earlier, the proposed lexicon contains words from 90 different domains. For example, মহাবিশ্ব/ universe, প্রকৃতি/ nature, পৃথিবী/earth, গাছপালা/flora, ইন্দ্রিয়–অনুভূতি/sense-perception, কাল/time, ঋতু/season, and বয়স/ age are different domains. Each domain is a collection of concepts, for example, সূর্য is a concept under the domain মহাবিশ্ব. Moreover, সূর্য also belongs to the domain of প্রকৃতি/ nature as well as পৃথিবী/earth connected using the primary links. Till date, there are all together there are 757 unique concepts under the head of 90 domains. These concepts are divided into sub-concepts in some cases. The sub-concepts do not have any specific name. We have provided each sub-concept a unique id. The words (mainly nouns, pronouns, adjectives, adjective-nouns and verbal adjectives) have been distributed into separate clusters attached to the concepts or sub-concepts and they form the leaves of the hierarchy. There is a common root node as antecedent to all the categories. Corresponding to each concepts, there are two types of clusters: one contains the exact synonyms and the clusters of the other type contain related words or attributes. The words belonging to the same cluster are synonymous. For example, consider the concept "সাহসিকতা". The lexical items under this concept like, সাহস , নির্ভীকতা ,নির্ভয়তা are all synonyms to each other. Therefore they form a cluster. On the other hand, lexical items like, দুঃসাহস ,ইচ্ছাশক্তি ,তেজ ,শৌর্য ,বীর্য are although semantically related but not exactly synonymous to সাহসিকতা therefore they form a separate cluster in the lexicon. Moreover every concept contains a set of antonyms associated with them. The antonyms are situated within the same clusters where synonyms are present. However, they are separated from the synonym through a specific *antonym marker*. For example, দেশদ্রোহী/traitor is under the concept স্বদেশ/native land in adjective section with [*/antonym*] tag. Every *category, concept and cluster* has distinct identification numbers which are stored along with the lexemes for further processing. An illustration of the hierarchical representation of domain→concepts→sub-concepts and→ clusters is shown in Figure 2.

In Figure 2, the category id of মহাবিশ্ব–প্রকৃতি–পৃথিবী–গাছপালা is 1, মহাবিশ্বhas sub-category id 1.1 meaning it is the 1st sub-category of category 1 and cluster id 1.1.3 as it belongs to the synonym cluster of 1.1. The member relations of words with their clusters have been shown in dashed lines and the round dotted line with arrowhead indicates that it is a primary link.

Each of the concepts has their corresponding synonyms and similar or related words. Different groups of words that are associated with a concept are organized together. Relevant information such as part of speech (POS) is corresponding to every group of words which are specified by tags like বি. /Noun, বিণ. /Adjective etc. If any word is mythological word then a tag like [পৌরা.] is specified before that word. Words having hyphen '-' at its end can be used as verb; e.g. ধরা–/catch, খেলা-/play etc. There are some words which are nouns or adjectives but we can use those words as verbs by appending some words with those. In our corpus the words to be appended are specified after the main word within

parenthesis like শিকার (ক). শিকার/Hunt is basically a noun but we can use it as a verb by appending করা/do to it which is indicated by (ক). There are several tags like this e.g. (দে) indicates দেওয়া, (হ) indicates {হওয়া} etc. In case of collective noun the collection of words are specified within square brackets separated by semicolon (;) after the word e.g. সপ্তপাতাল [অতল ; বিতল ; সুতল ; তলাতল ; মহাতল ; রসাতল ; পাতাল] , here সপ্তপাতাল is a collective noun .The most important thing in this corpus is that even if two words are orthographically, phonologically same but semantically different then no cross-reference occurs between them e.g. the word কলা means art; it is a type of fruit(banana) also; these two occurrences of word কলা do not have any cross links among each other as semantically these two senses are not close to each other.

4. MEASURING SEMANTIC SIMILARITY BETWEEN BANGLA WORDS

Many approaches have been taken to measure the semantic similarity between categories or words (described in section Background Study) such as information theoretic approaches, graph-based approaches. Here, we have proposed a simple graph based semantic similarity measure on our proposed lexicon. We have also verified it with user feedbacks.

Table 1: Edge Weight Distribution

Sr. No.	Type of link	Link weight (c is a constant whose value can be adjusted accordingly)
1.	**member** relation : between a cluster and a word under it	c
2.	**is-a** relation : between a sub-concept and cluster under it	$(c/2)+(c/(4*x))$
3.	**is-a** relation : between a concept and sub-concept under it	$(c/2)+((3*c)/(4*x))$
4.	**is-a** relation : between a concept and cluster under it	$c+(c/x)$
5.	**is-a** relation : between a category and concept under it	$c+(2*c/x)$
6.	**is-a** relation : between the root and category under it	$c+(3*c/x)$
7.	**primary link** : between a word and a concept(or a sub-concept)	$(3*c/2)+(c/(2*x))$
8.	**secondary link** : between a word and a concept(or a sub-concept	$(5*c/2)+(2*c/x)$

In our proposed lexical representation, the nodes on the top represents generalized concepts and as one goes down the hierarchy the nodes represent more specialized concepts. Therefore, the distance between a category and one of its concepts is greater than that between a concept and one of its clusters or sub-concepts (if exists). To capture this in our similarity measure we have assigned edge-weights to represent the relative distances. There are 8 types of link in this organization. The assigned weights of those links are described in details in table 1.

We have assumed that the all the nodes at a particular level are equal in weight. The semantic distance between any pair of words (w_i, w_j) is measured by the shortest path distance between them-

$$\boldsymbol{similarity\ score}(w_i, w_j) = x/\sum_{edge \in shortestpath(w_i,w_j)} \boldsymbol{edge\_weight} \quad \ldots (1)$$

In Equation (1) x is a constant signifying the scale of measurement. We have taken $c = 0.5$ and $x = 10$, so that a pair of synonyms has a score of 10 out of 10. The distribution of edge weights in the lexicon are shown in 3. Therefore, from Equation (1), the semantic similarity values between different types of word pairs are as depicted in Table 2.

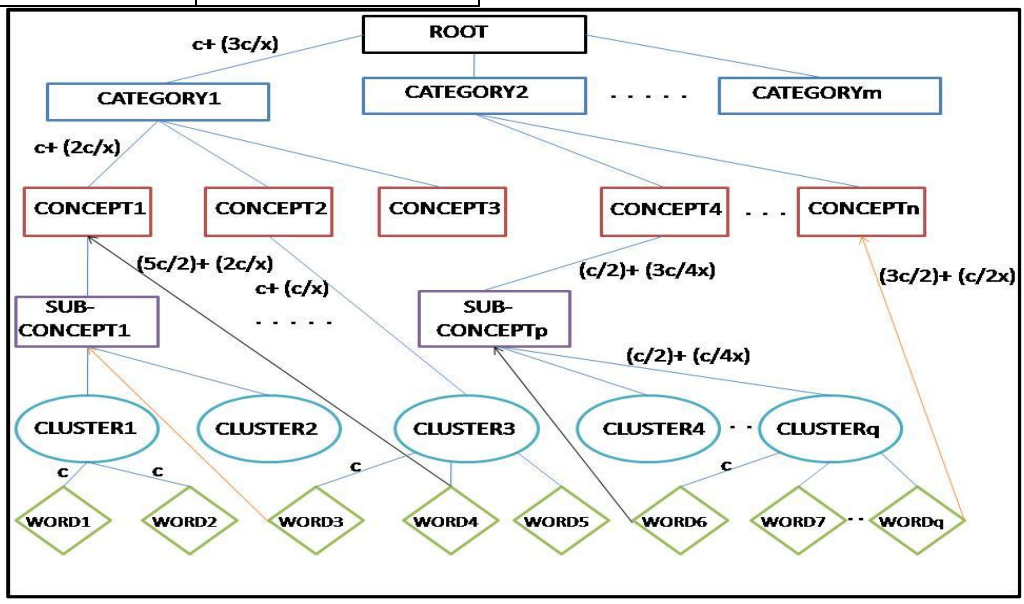

Figure 3: Edge-weight Distribution in Lexicon

Table 2: Similarity Scores

Case	Path to traverse	Score (in a scale of 10)($c = 0.5$)
both the words are in same cluster	$w_i \rightarrow cluster_{ij} \rightarrow w_j$	$x/(2 * c) = 10$
both the words are in same sub-concept, but in different clusters	$w_i \rightarrow cluster_i \rightarrow sub\_concept_{ij} \rightarrow cluster_j \rightarrow w_j$	$x/\big(2 * (c + (c/2) + (c/(4 * x)))\big) = 6.55$
both the words are in same concept , but in different clusters	$w_i \rightarrow cluster_i \rightarrow concept_{ij} \rightarrow cluster_j \rightarrow w_j$	$x/\big(2 * (c + c + (c/x))\big) = 4.76$
both the words are in same category , but different concepts	$w_i \rightarrow cluster_i \rightarrow concept_i \rightarrow category_{ij} \rightarrow concept_j \rightarrow cluster_j \rightarrow w_j$	$x/\big(2 * (c + c + (c/x) + c + (2 * c/x))\big) = 3.03$
both the words are from different categories	$w_i \rightarrow cluster_i \rightarrow concept_i \rightarrow category_i \rightarrow root \rightarrow category_j \rightarrow concept_j \rightarrow cluster_j \rightarrow w_j$	$x \,/\, ((2 * (c + c + (c \,/\, x) + c + (2 * c \,/\, x) + c + (3 * c \,/\, (x)))) = 2.17)$
both the words are from different concepts, but connected through primary link to sub-concept	$w_i \rightarrow sub\_concept_j \rightarrow cluster_j \rightarrow w_j$	$x \,/\, (((3 * c \,/\, 2) + (c \,/\, ((2 * x))) + (c \,/\, (2)) + (c \,/\, ((4 * x))) + c) = 6.5)$
both the words are from different concepts, but connected through primary link to concept	$w_i \rightarrow concept_j \rightarrow cluster_j \rightarrow w_j$	$x \,/\, (((3 * c \,/\, 2) + (c \,/\, ((2 * x))) + c + (c \,/\, (x)) + c) = 5.48)$
both the words are from different concepts, but connected by secondary link to sub-concept	$w_i \rightarrow sub\_concept_j \rightarrow cluster_j \rightarrow w_j$	$x \,/\, (((5 * c \,/\, 2) + (2 * c \,/\, x) + (c \,/\, 2) + (c \,/\, ((4 * x))) + c) = 4.73)$
Both the words are from different concepts, but connected by secondary link to concept	$w_i \rightarrow concept_j \rightarrow cluster_j \rightarrow w_j$	$x \,/\, (((5 * c \,/\, 2) + (2 * c \,/\, x) + c + (c \,/\, (x)) + c) = 4.16)$

5. EVALUATION

In order to evaluate our proposed semantic similarity measure, we have selected 400 different Bangla word pairs from our developed semantic lexicon. The selection of these word pairs were done in a pseudo-random manner. 300 word pairs were selected in a controlled manner. These word pairs were chosen from six different categories of relations in the following way:

- **Category 1**:50 pairs had both the words from the same cluster, i.e. synonyms.
- **Category 2**:50 pairs had words from different clusters of the same concept.
- **Category 3**:50 pairs had words from different concepts of the same category.
- **Category 4**:50 pairs had words from belonging from different categories.
- **Category 5**:50 pairs had words connected by primary links to concept.
- **Category 6**:50 pairs had words connected by secondary link to concept.

Another set of 100 word pairs were randomly chosen from the lexicon. These word pairs may or may not have any semantic relationship among them. We have also chosen another set of 200 word pairs which do not share any semantic relationship among them. Altogether, 600 Bangla word pairs were selected for our evaluation purpose.

60 different native speakers of Bangla participated in the experiment with age between 23 years to 36 years. All of them hold a graduate degree in their respective fields and 10 have a post graduate degree.

Each participant was provided the same set of 600 Bangla word pairs. The participants were asked to assign a score from 1 to 10 to each of the 300 word pairs based on their degree of relatedness: 1 for the lowest or no connectivity and 10 for the highest connectivity or synonyms.

5.1 Result and Discussion

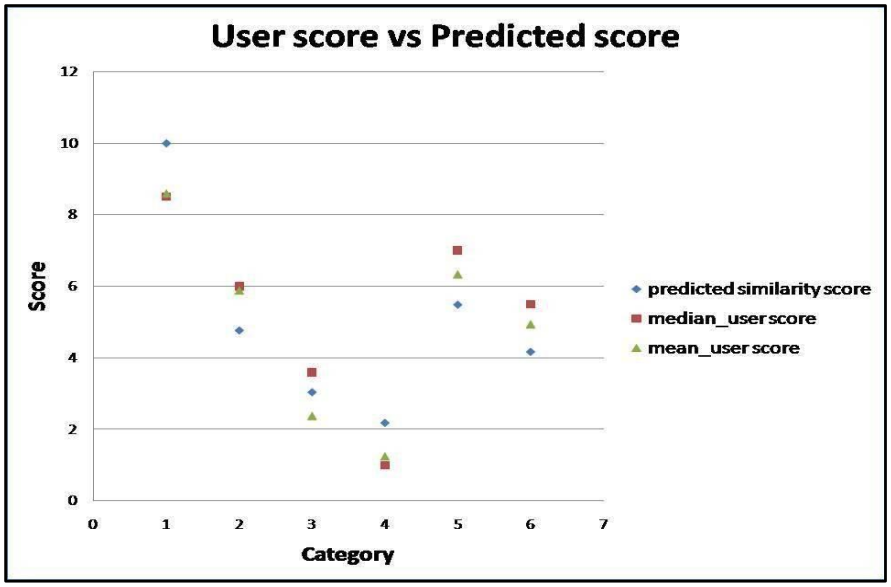

Figure 4: Performance analysis of user rating versus predicted measure

Perceiving semantic similarity or relatedness between a pair of words or concepts denoted by them depends on cognitive skill, domain or language knowledge and background of the user. Corresponding to each of the six types of words taken for user study, we have calculated both median and mean of user ratings. Mean has been used because of its popularity and common use, but as mean is very sensitive to outliers or extreme values median has also been taken into account. The table 3 below shows the outcomes of the user validation. Figure 4 below demonstrates the results graphically, it can be easily seen that the user ratings and our proposed measure are very close to each other.

Table 3: User Score versus Predicted Score

Category	1	2	3	4	5	6
Median User Rating	8.5	6	3.59	1	7	5.5
Mean User Rating	8.6	5.89	2.38	1.25	6.34	4.94
Predicted similarity score	10	4.76	3.03	2.17	5.48	4.16

One interesting point to be noted here is that the overall mean and median of user ratings for category 1 is less than 10. This means synonyms are not always perceived as exactly similar to each other. Spearman's rank correlation[7] of the predicted semantic similarity measure with the median values of user scores corresponding to each of the 50 word pairs is 1. To depict the subjectivity of users' perception, we have plotted the median values against our proposed scores (refer to section 5). As can be seen from the figure 5, there are few outliers in the dataset who have median values far from the group mean and median (type 1). Another type (type 2) of word pair is of interest as they have significant difference (greater than 1) between mean and median value, which implies that user ratings contain some extreme values. The pairs belonging to each type are given below in table 4.

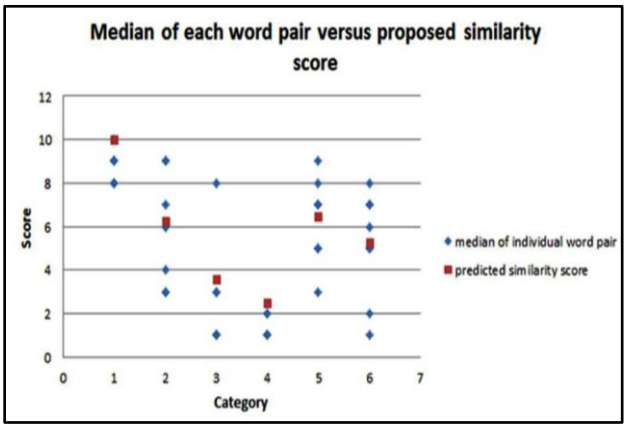

Figure 5: Comparisons of ratings of individual pairs with proposed scores

Table 4: List of Type1 and Type2 words

Category	Word-pair	Type
1	দুর্গা/Durga-ভগবতী/Bhagovati	2
2	রুচি/interests-রমণীয়/beautiful	2
3	বন্যা/flood-পর্বত/mountain	2
5	গ্রহজগৎ/planetary system-সৌরলোক/solar system	2
5	কৃষিজমি/farm land-ফসল/crop	2
2	নগ্নতা/naked-বিবস্ত্র/undressed	1
2	আলাদা/different-বিভেদ/discriminate	1
5	গমন/go, travel-যাওয়া/departure	1
5	শিলাবৃষ্টি/hail-	1

7 http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

	বরফপড়া/snowfall	
6	ভরাকোটাল/ high-tide- জলপ্লাবন/flood	1
3	সাফল্য/success- খ্যাতি/fame	1,2
6	হিমশৈল/iceberg-নুড়ি/pebbles	1,2
6	ক্রমশ/continued- মন্থরতা/slowness	1,2

As can be seen from the above table, word-pairs like (দুর্গা/Durga—ভগবতী/Bhagovati) demands a certain level of knowledge about the mythology to be perceived as synonyms, therefore, the user scores corresponding to this kind of word pairs also vary from person to person. Again, the similarity for the word pairs (গ্রহজগৎ/planetary system- সৌরলোক/solar system) and (কৃষিজমি/farm land-ফসল/crops) depend on how a user connects the two concepts in her cognition. The type 1 word pairs such as (নগ্নতা/naked- বিবস্ত্র/undressed), (শিলাবৃষ্টি/hail-বরফপড়া/snowfall) and (সাফল্য/success-খ্যাতি/fame) has been marked as synonyms or highly similar by the users. These phenomena demonstrate the confusion in distinguishing synonyms and very closely related concepts or words, especially those which are used alternatively in frequent situations. Three pairs belong to both types signifying they have been perceived as very close by most of the users and at the same time have got extreme values from the rest.

6. CONCLUSION AND FUTURE ASPECTS

We have proposed here a hierarchically organized semantic lexicon in Bangla and also a graph based edge-weighting approach to measure semantic similarity between two Bangla words. The similarity measures have been verified using user studies. We have also developed a graphical user interface to represent the lexical organization. Our proposed lexical structure contains only relations based on semantic association. We have included the frequency of each word over five Bangla corpuses in our lexical structure and also associated more details to words such as, whether the words are mythological or not, whether it can be used as verb or not, in order to use the word as a verb which word should be appended to it etc. As we have earlier discussed, this lexicon can be used in various applications like categorization, semantic web, and natural language processing applications like, document clustering, word sense disambiguation, machine translation, information retrieval, text comprehension and question-answering systems. We can also use it as a tool to improve the readability of text. For example we can substitute those words which are not understandable by reader with some easy words from the same cluster so that the sense of the sentence remain same. We can also use it as a tool to increase anyone's vocabulary.

In future, we will try to associate more details to words such as their pronunciations, distribution in spoken corpus, and word frequency history over time etc. We will tag specific relations between concepts or sub-concepts and clusters that mean how a cluster of words is related to a concept or sub-concept. We are thinking of annotating it manually. We will try to incorporate the information content of the words or other types of nodes in the similarity measure and subsequently verify them against user ratings as well as other automatic applications like text simplification, WSD etc. We still have to consider the relative difficulty of each word based on their corpus frequency or probability of occurrence. Also, all the clusters belonging to a common concept and all concepts descending from a common category have been assumed as equal. From the results of the user study it seems that there should be relative gradations of degree of similarity in these cases. We need to include these considerations in our measurement framework in order to achieve better correlation with users' cognitive perception. As evident from the users' feedbacks, the perception of semantic similarity between a pair of words varies largely according to user background. There should be an efficient mechanism to take into account the user's background. The present lexical structure contains static edges representing an ideal situation; a lexicon having dynamic connectivity can be helpful in understanding the effect of learning on the organization of mental lexicon.

7. REFERENCES

[1] Aitchison, J. (2012). Words in the mind: An introduction to the mental lexicon Wiley-Blackwell.

[2] Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). Adding dense, weighted connections to wordnet. In Proceedings of the Third International WordNet Conference, pages 29–36.

[3] Das, A. and Bandyopadhyay, S. (2010). Semanticnet-perception of human pragmatics. In Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon, pages 2–11, Beijing, China. Coling 2010 Organizing Committee.

[4] Fellbaum, C. (2010). Wordnet.Theory and Applications of Ontology: Computer Applications, pages 231–243.

[5] Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

[6] Kim, Y. and Kim, J. (1990). A model of knowledge based information retrieval with hierarchical concept graph. Journal of Documentation, 46(2):113–136.

[7] Lee, J., Kim, M., and Lee, Y. (1993). Information retrieval based on conceptual distance in is-a hierarchies. Journal of documentation, 49(2):188–207.

[8] Levelt, W. (1989). Speaking: from intention to articulationmit press. Cambridge, MA.

[9] Li, Y., Bandar, Z., and McLean, D. (2003).An approach for measuring semantic similarity between words using multiple information sources.Knowledge and Data Engineering, IEEE Transactions on, 15(4):871–882.

[10] Liu, H. and Singh, P. (2004).Conceptnet—a practical commonsense reasoning tool-kit.BT technology journal, 22(4):211–226.

[11] Mukhopadhyay, A. (2005). Samsad Samarthaksabdokosh. SahityaSamsad, 12 edition.

[12] Muller, S. (2008).The mental lexicon. GRIN Verlag.

[13] Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989).Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on, 19(1):17–30.

[14] Resnik, P. (1993a). Selection and information: a class-based approach to lexical relationships. IRCS Technical Reports Series, page 200.

[15] Resnik, P. (1993b). Semantic classes and syntactic ambiguity. In Proc. of ARPA Workshop on Human Language Technology, pages 278–283.

[16] Richardson, R., Smeaton, A., and Murphy, J. (1994).Using wordnet as a knowledge base for measuring semantic similarity between words.Technical report, Technical Report Working Paper CA-1294, School of Computer Applications, Dublin City University.

[17] Roy, M. and Muqtadir, M. (2008).Semi-automatic building of wordnet for Bangla.PhD thesis, School of Engineering and Computer Science (SECS), BRAC University.

[18] Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., and Scheffczyk, J. (2010).Framenet ii: Extended theory and practice, available online at h ttp. framenet. icsi. berkeley. edu.

[19] Seashore, R. and Eckerson, L. (1940). The measurement of individual differences in general English vocabularies. Journal of Educational Psychology; Journal of Educational Psychology, 31(1):14.

[20] Tversky, A. (1977). Features of similarity. Psychological review, 84(4):327.

[21] Wang, T. and Hirst, G. (2011).Refining the notions of depth and density in wordnet-based semantic similarity measures. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

[22] Biemann, C.(2007).Unsupervised Natural Language Processing. In Proceedings of the NAACL-HLT 2007 Doctoral consotorium,Rochester,April 2007, pages 37-40.

[23] Lin, D.(1998).Automatic Retrieval and Clustering of Similar Words. In COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 2,Pages 768-774.

[24] Biemann, C., Shin,S., Choi,K.(2004).Semiautomatic extension of CoreNet using a bootstrapping mechanism on corpus-based co-occurrences. In COLING '04 Proceedings of the 20th international conference on Computational Linguistics,Article No. 1227.

[25] Davidov, D., Rappoport, A.(2006).Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words.In ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics,Pages 297-304.

[26] Davidov, D., Rappoport, A., Koppel, M.(2007).Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,Prague,Czech Republic,June 2007,pages 232-239.

[27] Biemann, C.(2006).Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In TextGraphs-1 Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing , Pages 73-80.

[28] Gentner, D.(1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj, editor, Language development: Vol. 2. Language, thought, and culture, pages 301-334. Erl-baum, Hillsdale, NJ.

[29] Quasthoff,U.,Biemann, C.,Wolff, C.Named entity learning and verification: expectation maximization in large corpora,COLING-02 proceedings of the 6th conference on Natural language learning - Volume 20Pages 1-7.

[30] Sinha, Manish and Reddy, Mahesh and Bhattacharyya, Pushpak, (2006) "An approach towards construction and application of multilingual indo-wordnet", 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea.