A New Semantic Lexicon and Similarity Measure in Bangla

Manjira Sinha, Abhik Jana, Tirthankar Dasgupta, Anupam Basu
Indian Institute of Technology Kharagpur
{manjira87, abhikjana1, iamtirthankar, anupambas}@gmail.com

ABSTRACT

The *Mental Lexicon* (ML) refers to the organization of lexical entries of a language in the human mind.A clear knowledge of the structure of ML will help us to understand how the human brain processes language. The knowledge of semantic association among the words in ML is essential to many applications. Although, there are works on the representation of lexical entries based on their semantic association in the form of a lexicon in English and other languages, such works of Bangla is in a nascent stage. In this paper, we have proposed a distinct lexical organization based on semantic association between Bangla words which can be accessed efficiently by different applications. We have developed a novel approach of measuring the semantic similarity between words and verified it against user study. Further, a GUI has been designed for easy and efficient access.

KEYWORDS : Bangla Lexicon, Synset, Semantic Similarity, Hierarchical Graph

1 Introduction

The *lexicon* of a language is a collection of lexical entries consisting of information regarding words and expressions, comprising both *form* and *meaning* (Levelt,). *Form* refers to the orthography, phonology and morphology of the lexical item and *Meaning* refers to its syntactic and semantic information.

The term *Mental Lexicon* refers to the organization and interaction of lexical entries of a language in the human mind. Depending on the definition of *word*, an adult knows and uses around 40000 to 150000 words. Yet, it has been estimated that an adult can recognize a *word* in her native language in less than 200ms and can reject a non-word in less than 500ms (Aitchison, 2012; Muller, 2008; Seashore and Eckerson, 1940). Therefore, the storage and retrieval mechanisms of the brain have to be efficient enough to facilitate such super-fast access. Words in the mental lexicon are assumed to be associated at various levels of linguistic features such as, *orthography, phonology, morphology* and *semantics*. Although a vast amount of research is going on the mental lexicon, the precise natures of the relations are yet to be explored. The knowledge of semantic association among the words in mental lexicon is essential to many areas such as, developing pedagogical strategies, categorization, semantic web, natural language processing applications like, document clustering, word sense disambiguation, machine translation, information retrieval, text comprehension, and question-answering systems, where the perception of the target user group plays an important role.. However, as we cannot 'look into the mind' to know the exact structure of the mental lexicon, we try to simulate its behaviour with the help of external models.

The rich repertoire of literature on the structure, organization and representation of lexical entries includes simple organization schemes like Dictionary and Thesaurus to more complex ones like

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pages 171–182,
COLING 2012, Mumbai, December 2012.

WordNet (Fellbaum, 2010) and ConceptNet (Liu and Singh, 2004) and also methods to measurethe degree of semantic similarity among the lexemes.

Bangla is an Indo-Aryan language having about 193 million native and about 230 million total speakers. Despite being so popular, very few attempts have (Roy and Muqtadir, 2008; Das and Bandyopadhyay, 2010) been made to build a semantically organized lexicon of substantial size in Bangla. Hence, we propose a distinct lexical organization.

The objective of this work is to design and develop a Bangla lexicon based on semantic similarity among Bangla words, which is suitable of automatic access mechanismsand can be used further in various applications like as mentioned above. The design is based on the *Samsad Samarthasabdokosh* (Mukhopadhyay, 2005). The lexicon is hierarchically organized and divided according to the categories or domains represented by different segments. The categories are further divided into sub-categories. The words are grouped into clusters along with their synonyms. Weighted edges between different types of words related to same or different concepts or categories exist, denoting the semantic distance between them. We have also developed a Graphical User Interface on top of the lexicon, which can be used for efficient and easy access. This is an on-going project with an aim of creating an organization containing 50,000 words.

The organization of the paper is as follows: section 2 contains the related works; we have also pointed out some of the differences of our proposed structure with WordNet in section 2; section 3 explains the construction of the lexicon and the GUI; section 4, describes the proposed approach of predicting semantic similarity between words; in section 5 we have discussed the user study; conclusions and future thoughts have been included in the last section.

2 Related work

A number of works have been done semantic relation based representations include simple organizational schemes like Dictionary and Thesaurus to more complex ones like WordNet (Fellbaum, 2010) and ConceptNet (Liu and Singh, 2004) and others (Ruppenhofer et al., 2010).Words in WordNet are organized around semantic groupings called *synsets*. Each synset consists of a list of synonymous word forms and semantic pointers that describe relationships among the synsets.However, WordNet suffers from several limitations(Boyd-Graber et al., 2006).ConceptNet is a semantic network containing different types of *concepts* and relationships among them. Here, concepts are represented by words or short phrases and relationships can be of many kinds such as, *MotivatedByGoal*, *UsedFor*, *can cross.*

According to the most recent reference to a Bangla WordNet (Roy and Muqtadir, 2008), 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. It contains around 639 synsets and 1,455 words[1]. The assumptions that have been taken are: Bangla and English have significant amount of linguistic similarities and Bangla word senses can be clearly justified by a Bangla-English-Bangla dictionary.

Our proposed lexical representation is different from WordNet in many respects. Some of the important differences being:

[1] http://bn.asianwordnet.org/

- No cross parts of speech links are there in the WordNet. That means no link between an entity and its attributes.
- Several lexical and semantic relations are not included in the WordNet such as "actor"([book]-[writer]), "instrument"([knife]-[cut]), but these are perceived as related by human cognition. In our framework these types of relations are, for example under the node [book], [writer] is there in [noun-adjective] type of cluster. [Knife], [cut] are also under same node [weapon] but in different clusters. These kinds of relations can be helpful in word sense disambiguation applications.
- Relational links are qualitative rather than quantitative in WordNet.. In our system we have given weight on different type of links keeping in mind the semantic closeness of the nodes they connect. Moreover, in our structure, there exists a path between each possible word pair.

Our proposed semantic similarity based lexical organization is not a substitution of WordNet; rather it tries to address some of the aspects which are still not incorporated in the WordNet framework. It is useful especially in case of a resource poor language like Bangla.

2.1 Work on measuring semantic similarity among words

There exist many approaches to measure semantic similarity between words; some of them are discussed here. Tversky's feature based similarity model (Tversky, 1977), is among the early works in this field. Some scholars (Rada et al., 1989; Kim and Kim, 1990; Lee et al., 1993) have proposed the conceptual distance approach that uses edge weights, between adjacent nodes in a graph as an estimator of semantic similarity. Resnick (Resnik, 1993a; Resnik, 1993b) have proposed the information theoretic approach to measure semantic similarity between two words. Richardson et. al. (1994) has proposed an edge-weight based scheme for Hierarchical Conceptual Graphs (HCG) to measure semantic similarity between words. Efforts (Jiang and Conrath, 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 (Li et al., 2003). Wang and Hirst (2011) have criticized the traditional notions of the depth and density in a lexical taxonomy. However, almost all of the attempts described above have been taken in English based on the representation of WordNet. Das and Bandopadhaya (2010) have proposed a SemanticNet in Bangla, where the relations are based on human pragmatics.

3 Construction of the Proposed Lexicon

We have taken the *Samsad Samarthasabdokosh*by Ashok Mukhopadhyay(2005) as the basis for our proposed lexical representation in Bangla. The book contains 757 main words distributed in 30 different sections. Each section addresses a particular domain such as universe-nature-earth, life-living being-body etc. The main words have their corresponding synonyms and similar or related words. Different groups of words that are associated with a single main word are organized together. Relevant information such as Part-Of-Speech (POS) corresponding to every word and antonyms for adjectives are also mentioned. Two types of cross-references are present: one relates two main words or a single word which is simultaneously synonymous to two different main words and the other denotes multiple occurrences of the same. We have termed them as primary link and secondary link respectively. We have also analysed Bangla corpuses:

complete novel and story collection of Rabindranath Tagore, Bankimchandra Chattaopadhayay[2], collection of Bangla blogs over the internet, Bangla corpus by CIIL[3]Mysore and Anandabazar news corpus[4] and have prepared a list of around 4 lakh distinct words in Bangla with their corpus frequencies.

In order to build-up a semantic relation based lexical representation Bangla; we have constructed a hierarchical conceptual graph based on the above mentioned book. 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 frequency over the five mentioned corpuses, its association 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). The details of the organizational methodology are described below.

The 30 different sections have been considered as 30 root categories. Each category is a collection of concepts, e.g. ইন্দ্রিয়–অনুভূতি/sense-perception. 757 main words have been organized under the root categories as sub-categories, which are actually concepts, e.g. গন্ধ/smell.The words (mainly nouns, adjectives, verbal nouns and verbal adjectives) have been distributed into separate clusters attached to the sub-categories and they form the leaves of the hierarchy. There is a common root node as antecedent to all the categories. Corresponding to each sub-category, 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. Every category, sub-category and cluster has distinct identification numbers.

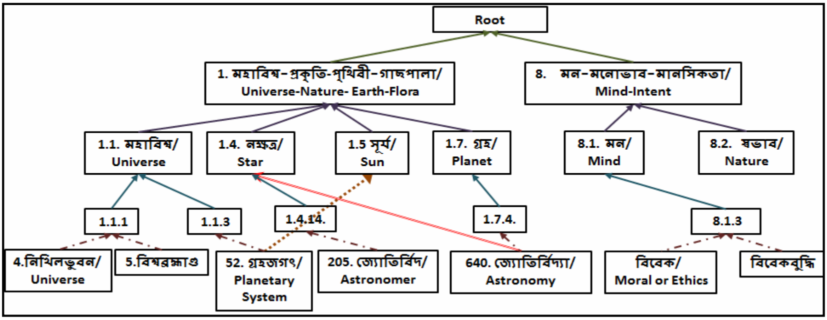

FIGURE 1- Partial view of our proposed lexicon

[2] http://www.nltr.org/

[3]http://www.ciil.org/

In figure 1, 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.

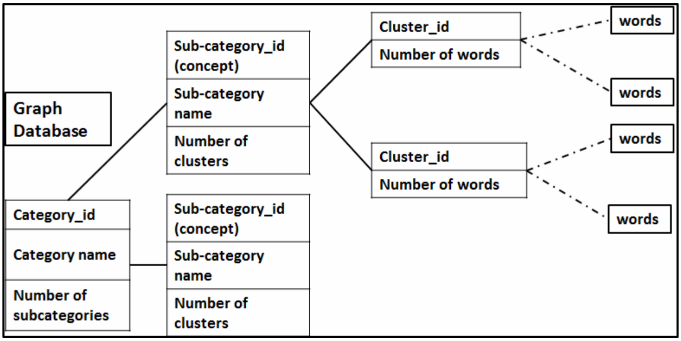

FIGURE 2-Simplified view of the underlying storage structure

Every word has been assigned an information array with 15 fields. They are:

- Serial_no: denotes the serial number of a word in the database
- Part-Of-Speech (POS)
- Corpus frequency
- Cluster_number: number of the cluster of the word
- SC_no.: number of the sub-category under which the word belongs
- SSC_no: number of the sub-sub-category of a word (applicable to few words)
- C_id: number of the category under which the word resides
- P_link: pointer to the cluster id under the sub-category specified by the primary link
- S_link: pointer to the cluster id under the sub-category specified by the secondary link
- Antonym: cluster id of the antonym(s) of the word
- Myth: a flag to indicate any mythical relation to the word
- Details: Serial no. of all the words in the collection denoted by the present word (if it is a collective noun)
- G_word: a pointer to the general word denoting the collection in which the present word belongs
- Verb: a flag to indicate whether the word can be also used as a verb or not.
- To_verb: contains the word which can be appended to the present word to make it possible to be used as a verb.

The fields from 1 to 4 and 5 are available for every word; rests of the fields have values, if available or they have been assigned null. Words belonging to multiple clusters have more than one 14-field information vector associated with them. Refer to the examples below for details:

Word	বাঁধ/dam	স্বর্গ/heaven	চতুর্দশভুবন/fourteen worlds	সমুদ্রযাত্রা/seatravel	আত্মঘাতী/sucidal	গ্রহজগৎ/planetary system	
Serial_no	1659	67	69	1440	6032	52	
Pos	বিশেষ্য [noun]	বিশেষ্য	বিশেষ্য	বিশেষ্য	বিশেষণ[adjective]	বিশেষ্য	
Corpus frequency	185	442	-	27	1069	-	-
Cluster_no	7	7	7	8	56	3	4
Sc_no	17	1	1	14	9	1	7
C_no	1	1	1	1	2	1	1
P_link	null	23.22.1	null	null	null	1.5.2	
S_link	1.47.22	null	null	null	null	null	
Antonym	null	null	null	null	2.58.58	null	
Myth	null	1	1	null	null	null	
Details	null	null	71,72	null	null	null	
G_word	null	null	null	null	null	null	
Verb	null	null	null	null	null	null	
M_to _verb	null	null	null	(ক)	null	null	

TABLE 1- organization of word database

3.1 Graphical User Interface

We have also developed a Graphical User Interface based on the lexical representation described above. It can perform two jobs. First, it can be used to find the details about a particular word or category present in the database. A user can provide input in two different ways: directly typing the word or selecting from the list of words of different parts of speeches. For the ease of typing Bangla, we have also provided a Bangla virtual keyboard associated with the GUI. Given a word, the system outputs all the available fields associated with the word. It also provides the name and link of the corresponding sub-category and category so that the user can view details about just by clicking on them. If a word belongs to more than one cluster or part-of speech, the GUI shows all the associated clusters and sub-concepts, concepts. User can also navigate to the sub-concept(s) associated by primary link or secondary link with the help of the GUI. Second, given two words as input the GUI also calculates the degree of semantic similarity between them along with their corresponding positions in the lexical representation. The method of obtaining the semantic similarity or relatedness measure has been described in the next section.

4 Semantic Similarity Measure between Bangla Words

As we have discussed in the above sections, along with relating words semantically, the mental lexicon also assigns a degree of similarity between them. Here, we have proposed a simple graph based semantic similarity measure on our proposed lexicon. We have also verified it with user feedbacks. In our proposed lexicon, the nodes from the top to bottom represent generalized to more specialized concepts. Therefore, the semantic distance or edge weights decrease as one moves down the hierarchy. There are 8 types of direct link in the organization:

Sr. No.	Type of link	**Link weight (c is a constant whose value can be adjusted according to the need)**
1.	**member relation**: between a word and its cluster	c
2.	between a cluster and its sub-category	$\frac{c}{2}+\frac{c}{x}$
3.	between a cluster and its sub-sub-category (if present)	$\frac{c}{2}+0.5*\frac{c}{x}$
4.	**is-a** relation: between a sub-sub-category and its sub-category (if present)	$c+\frac{c}{x}$
5.	**is-a** relation: between a sub-category and its root category.	$c+\frac{2c}{x}$
6.	between a category and the root.	$c+\frac{3c}{x}$
7.	**primary_link**: between a word and a sub-category (according to the representation, this distance is greater than a member relation but lesser than the total path length between word and its sub-category)	$c+\frac{c}{2}$
8.	**secondary_link**: between a word and a sub-category (this distance is greater than the distance between a sub-category and its category)	$2c+\frac{2c}{x}$

TABLE 2- Edge-weight distributions

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:

$$similarity\ score(w_i, w_j) = {}^{x}\!\big/\textstyle\sum_{i \in shortest\ path(w_i w_j)}(edge_i - weight) \quad ..... (1)$$

Here, 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.Therefore, from table 2and equation (1), the semantic similarity values between different types of word pairs are as shown in table 3.

In order to verify whether the proposed approach to measure semantic similarity or relatedness between a pair of words can actually represent the degree of similarity as perceived by human cognition, we have carried out a user survey. The details of the study have been described in the next section.

Case	**Score (in a scale of 10)**
both the words are in same cluster (synonym)	$x\ /2*c = 10$
both the words are in same sub-category (S_i), but in different clusters	$x/(c + 2(c/2 + c/x) + c) = 6.25$
both the words are in same category (C_i), but different sub-categorys	${}^{x}/_{2}\left(c + \left(\frac{c}{2} + {}^{c}/_{x}\right) + (c + 2\,{}^{c}/_{x})\right) = 3.57$
both the words are from different categorys	${}^{x}/_{2}\left(c + \left(\frac{c}{2} + {}^{c}/_{x}\right) + \left(c + 2\,{}^{c}/_{x} + (c + 3\,{}^{c}/_{x})\right)\right) =$ 2.5
both the words are from different sub-categorys, but connected through primary_link	${}^{x}/\left(\frac{3c}{2} + \left(\frac{c}{2} + {}^{c}/_{x}\right) + c\right) = 6.45$
both the words are from different sub-categorys, but connected by secondary_link	${}^{x}/\left(\left(2c + {}^{2c}/_{x}\right) + \left(\frac{c}{2} + {}^{c}/_{x}\right) + c\right) = 5.26$
Antonym is a special type of relation	-1

TABLE 3-Similarity scores

5 User Study

Participants: 25 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.

Experiment data selection and procedure: 50 word pairs were selected from the lexical representation. The word pairs were chosen from the six different categories of relations described in table 4 above, except antonyms. Each user was asked to assign a score from 1 to 10 to each of the 50 word pairs based on their degree of semantic relatedness: 1 for the lowest or no connectivity and 10 for the highest connectivity or synonyms.

5.1 Result and Discussions

Perceiving semantic similarity or relatedness between a pair of words or concepts denoted by them depends on the 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 outlier or extreme values median has also been taken into account. The table 4 below shows the outcomes of the user validation:

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	6.25	3.57	2.5	6.45	5.26

TABLE 4-User score versus predicted score

The figure 3 below demonstrates the results graphically, it can be easily seen that the user ratings and our proposed measure are very close to each other. 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[5] of the predicted semantic similarity measure with the median values of user scores corresponding to each of the 50 word pairs is 0.8. To depict the subjectivity of user's perception, we have plotted the median values against our proposed scores (refer to figure 4).

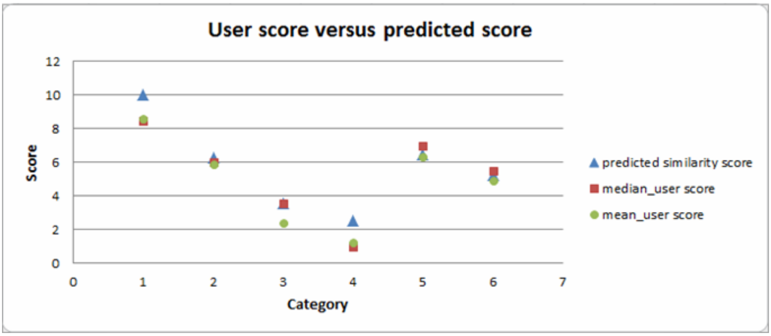

FIGURE 3-Performance analysis of user rating versus predicted measure ´

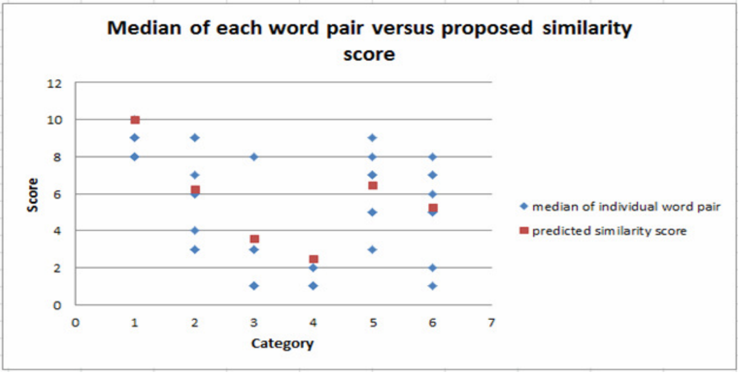

FIGURE 4- Comparison of ratings of individual pairs with proposed score

[5] http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

Figure 4 shows few outliers in the dataset that 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 values, which implies that user ratings contain some extreme values. The pairs belonging to each type are:

C	word pair	Type
1	দুর্গা--ভগবতী	2
2	রুচি/interests—রমণীয়/beautiful	2
3	বন্যা/flood—পর্বত/mountain	2
5	গ্রহজগ▯/planetary system—সৌরলোক/solar system	2
5	কৃষিজমি/farm land—ফসল/crop	2
2	নগ্নতা/naked—বিবস্ত্র/undresses	1

C	word pair	Type
2	আলাদা/different—বিভেদ/discriminate	1
5	গমন/go, travel—যাওয়া/departure	1
5	শিলাবৃষ্টি/hail -বরফপড়া/snowfall	1
6	ভরাকোটাল/hightide—জলপ্লাবন/flood	1
3	সাফল্য/success—খ্যাতি/fame	1, 2
6	হিমশৈল/iceberg—নুড়ি/pebbles	1, 2
6	ক্রমশ--মন্থরতা	1, 2

TABLE 5- List of type 1 and type 2 words. “C” implies Category.

As can be seen from the above table, word-pairs like (দুর্গা—ভগবতী) 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 (গ্রহজগ▯/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) have 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.

Conclusion and perspective

In this paper, we have proposed a hierarchically organized semantic lexicon in Bangla and also a graph based edge-weighting approach to measure the semantic similarity between two words. The similarity measures have been verified using user studies. We have included the frequency of each word over five Bangla corpuses in our lexical structure and also working on associating more details to words such as, their pronunciations, distribution in spoken corpus, word frequency history over time etc. Our proposed lexical structure contains only relations based on semantic association; we plan to extend the work to incorporate other kinds of relationships such as orthography, phonology and morphology to represent the human cognition more accurately.

Acknowledgements

We are thankful to Society for Natural Language Technology Research Kolkata for helping us to develop the lexical resource. We are also thankful to those subjects who spend their time to manually evaluate our semantic similarity measure.

References

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

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.

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.

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

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

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

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.

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

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.

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

Mukhopadhyay, A. (2005). *SamsadSamarthasabdokosh*. SahityaSamsad, 12 edition.

Müller, S. (2008).*The mental lexicon*. GRIN Verlag.

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.

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

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

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.

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

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*.

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.

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

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.