Can We Mimic Human Pragmatics Knowledge into Computational Lexicon?

Amitava Das
Department of Computer Science and Engineering
Jadavpur University
Jadavpur, Kolkata 700032, India
amitava.santu@gmail.com

Abstract

So far Natural Language Processing (NLP) research patronized much of manually augmented lexicon resources such as WordNet. But the small set of semantic relations like Hypernym, Holonym, Meronym and Synonym etc are very narrow to capture the wide variations human pragmatics knowledge i.e. a news article containing the themes, "Iraq", "Al-Qaeda", "9/11" and "Osama Bin Laden" might suggest the topic related to "terrorism". But no such information could be retrieved from available lexicon resources. SemanticNet is the attempt to capture wide context dependent semantic inference among various themes which human being perceives in their pragmatic knowledge, learnt by day to day cognitive interactions with physical world surrounded by them. One such effort towards capturing the human common sense is ConceptNet (Liu and Singh, 2004) for English. We extend our vision over the human common sense to human pragmatics and proposed semantic relations for every pair of lexicons cannot be defined by fixed number of certain semantic relation labels and thus we formulated a probabilistic score based technique. SemanticNet is a semantic network of lexicons to hold human pragmatic knowledge. Contextual semantic affinity inference in SemanticNet could be calculated by network distance. SemanticNet presently developed for Bengali language.

1 Motivation

Semantics (from Greek "*σημαντικός*" - *semantikos*) is the study of meaning, usually in language. The word "*semantics*" itself denotes a range of ideas, from the popular to the highly technical. It is often used in ordinary language to denote a problem of understanding that comes down to word selection or connotation. We studied with various Psycholinguistics experiments to understand how human natural intelligence helps to understand general semantic from nature. Our study was to understand the human psychology about semantics beyond language. We were haunting for the intellectual structure of the psychological and neurobiological factors that enable humans to acquire, use, comprehend and produce natural languages. Let's come with an example of conversation about movie between two persons.

Person A: Have you seen the movie '*No Man's Land*'? How is it?
Person B: Although it is good but you should see '*The Hurt Locker*'?

May be the conversation looks very casual, but our intension was to find out the direction of the decision logic on the Person B's brain. We start digging to find out the nature of human intelligent thinking. A prolonged discussion with Person B reveals that the decision logic path to recommend a good movie was as the Figure 1. The highlighted red paths are the shortest semantic thinking path.

We call it semantic thinking. Although the derivational path of semantic thinking is not such easy as we portrait in Figure 1 but we keep it easier for understandability. Actually a human try to figure out the closest semantic affinity node into his pragmatics knowledge by natural intelligence. In the previous example Person B find out with his intelligence that *No Man's Land* is a war movie and got Oscar award. Oscar award generally cracked by Hollywood movies and thus Person B start searching his pragmatics network to find out a

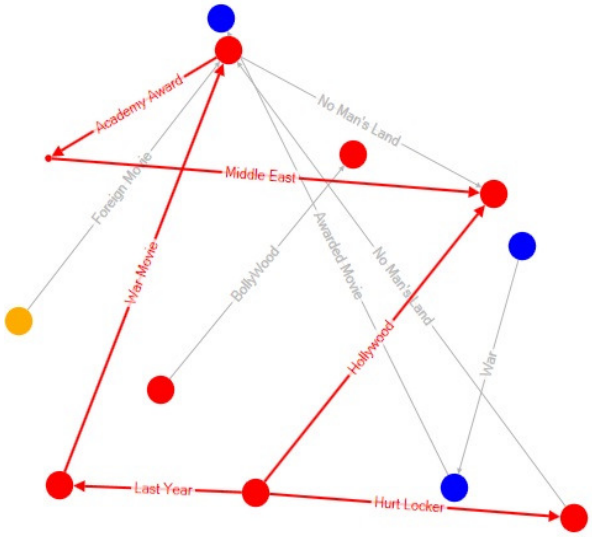

Figure 1: Semantic Thinking

movie fall into war genre, from Hollywood and may be got Oscar award. Person B finds out the name of a movie The Hurt Locker at nearer distance into his pragmatics knowledge network which is an optimized recommendation that satisfy all the criteria. Noticeably Person B didn't choice the other paths like Bollywood, Foreign movie etc. And thus our aim was to develop a computational lexicon structure for semantics as human pragmatics knowledge. We spare long time to find out the most robust structure to represent pragmatics knowledge properly and it should be easy understandable for next level of search and usability.

We look into literature that probably direct to the direction of our ideological thinking. We found that in the year of 1996 Push Singh and Marvin developed lexicon resources like ConceptNet (Liu and Singh, 2004). **ConceptNet-** ConceptNet is a large-scale semantic network (over 1.6 million links) relating a wide variety of ordinary objects, events, places, actions, and goals by 20 different link types, mined from the OMCS corpus.

ConceptNet is a nice path breaking research effort. But presently we extended our vision. One example may illustrate our idea. During psycholinguistics experiments we used to give various words to different people to understand human intelligence regarding language and lexicon. Suppose the word "*terminal*" may give an idea about many things like: aero planes, bus terminal, train terminal, space terminal etc. Undoubtedly it is a sense disambiguation issue. But a human can do it easily. If the word "*terminal*" comes with context "*Stephen Spielberg*" then it is definitely a movie. Therefore human intelligence work with prior bag-of-words concept and driven by domain prior knowledge (pragmatics). ConceptNet only provide word-word relation but a word has many-to-many relationship with multiple domains. A word "*terminal*" has many relationships with multiple domains.

In the last decade a revolutionary invention explain the question of how human understand languages. The biologists have discovered the human gene for languages and name it as FOXP2[1] (Vargha-Khadem et al, 1995). This invention may answer multiple questions raised by language researchers, NLP researchers and cognitive science peoples. The inventors of FOXP2 explained that how human being learns by their five sensible organs from nature and store it with the help of the innate gene in their intelligence structure. Let come with an example: It is generally seen any child utter "*fire*" after touching any hot object. It is due to their lack of lexicon knowledge. Actually during learning period, a human map unknown semantics to their known semantics. This process also helps during second language learning. But present computational technology unable to hold human basic senses and even medical science have no clear idea about how many basic senses a human being have.

[1] http://en.wikipedia.org/wiki/FOXP2
http://www.foxp2.org/

Therefore nothing could be done but only to rely on lexicon level. We experimented with available NLP techniques to capture human pragmatics to develop a computational lexicon, called SemanticNet.

The present task of developing SemanticNet is to capture semantic affinity knowledge of human pragmatics as a lexicon database. We extend our vision from the human common sense to human pragmatics and have proposed semantic relations for every pair of lexemes that cannot be defined by fixed number of certain semantic relation labels. Contextual semantic affinity inference in SemanticNet could be calculated by network distance and represented as a probabilistic score. SemanticNet is being presently developed for Bengali language.

2 Corpus

Present SemanticNet has been developed for Bengali language. Resource acquisition is one of the most challenging obstacles to work with electronically resource constrained languages like Bengali. Although Bengali is the sixth[2] popular language in the World, second in India and the national language in Bangladesh.

There was another issue drive us long way to find out the proper corpus for the development of SemanticNet. As the notion is to capture and store human pragmatic knowledge so the hypothesis was chosen corpus should not be biased towards any specific domain knowledge as human pragmatic knowledge is not constricted to any domain rather it has a wide spread range over anything related to universe and life on earth. Additionally it must be larger in size to cover mostly available general concepts related to any topic. After a detail analysis we decided it is better to choose NEWS corpus as various domains knowledge like Politics, Sports, Entertainment, Social Issues, Science, Arts and Culture, Tourism, Advertisement, TV schedule, Tender, Comics and Weather etc are could be found only in NEWS corpus.

Fortunately such corpus development could be found in (Ekbal and Bandyopadhyay, 2008) for Bengali. We obtained the corpus from the authors. The Bengali NEWS corpus consisted of consecutive 4 years of NEWS stories with various sub domains as reported above. For the present task we have used the Bengali NEWS corpus, developed from the archive of a leading Bengali NEWS paper[3] available on the Web. The NEWS corpus is quite larger in size as reported in Table 1.

Statistics	NEWS
Total no. of news documents in the corpus	108,305
Total no. of sentences in the corpus	2,822,737
Average no. of sentences in a document	27
Total no. of wordforms in the corpus	33,836,736
Avg. no. of wordforms in a document	313
Total no. of distinct wordforms in the corpus	467,858

Table 1: Bengali Corpus Statistics

2.1 Annotation

From the collected document set 200 documents have been chosen randomly for the annotation task. Three annotators (Mr. X, Mr. Y and Mr. Z) participated in the present task. Annotators were asked to annotate the theme words (topical expressions) which best describe the topical snapshot of the document.

The agreement of annotations among three annotators has been evaluated. The agreements of tag values at theme words level is reported in Table 2.

Annotators	X vs. Y	X Vs. Z	Y Vs. Z	Average
Percentage	82.64%	71.78%	80.47%	78.30%
All Agree	75.45%			

Table 2: Agreement of annotators at theme words level

3 Theme Identification

Term Frequency (TF) plays a crucial role to identify document relevance in Topic-Based Information Retrieval. The motivation behind developing Theme detection technique is that in many documents relevant words may not occur frequently or irrelevant words may occur frequently. Moreover for the lexicon affinity inference topic or theme words is the only strong clue to start with. The Theme detection technique has been proposed to resolve these issues to identify discourse level most relevant thematic nodes in terms of word or lexicon using a standard machine learning technique. The machine learning technique used here is Conditional Random Field (CRF). The theme word detection has been defined as a sequence labeling

[2] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[3] http://www.anandabazar.com/

problem using various useful depending features. Depending upon the series of input features, each word is tagged as either Theme Word (TW) or Other (O).

4 Feature Organization

The set of features used in the present task have been categorized as Lexico-Syntactic, Syntactic and Discourse level features. These are listed in the Table 3 below and have been described in the subsequent subsections.

Types	Features
Lexico-Syntactic	POS
	Frequency
	Stemming
Syntactic	Chunk Label
	Dependency Parsing Depth
Discourse Level	Title of the Document
	First Paragraph
	Term Distribution
	Collocation

Table 3: Features

4.1 Lexico-Syntactic Features

4.1.1 Part of Speech (POS)

It has been shown by (Das and Bandyopadhyay, 2009), that theme bearing words in sentences are mainly adjective, adverb, noun and verbs as other POS categories like pronoun, preposition, conjunct, article etc have no relevance towards thematic semantic of any document. The detail of the POS tagging system chosen for the present task could be found in (Das and Bandyopadhyay 2009).

4.1.2 Frequency

Frequency always plays a crucial role in identifying the importance of a word in the document or corpus. The system generates four separate high frequent word lists after function words are removed for four POS categories: adjective, adverb, verb and noun. Word frequency values are then effectively used as a crucial feature in the Theme Detection technique.

4.1.3 Stemming

Several words in a sentence that carry thematic information may be present in inflected forms. Due to non availability of good stemmers in Indian languages especially in Bengali, a stemmer based on stemming cluster technique has been used as described in (Das and Bandyopadhyay, 2010).

4.2 Syntactic Features

4.2.1 Chunk Label

We found that Chunk level information is very much effective to identify lexicon inference affinity. As an example:

(সত্যজিত রায়ের)/NP (মুক্তিপ্রাপ্ত ছবিগুলি)/NP (অনন্য)/NP (সাধারন)/JJP (।)/SYM

The movies released by Satyajit Roy are excellent.

In the above example two lexicons মুক্তি/release and ছবি/movie are collocated in a chunk and they are very much semantically neighboring in human pragmatic knowledge. Chunk feature effectively used in supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. In the task of identification of Theme expressions, chunk label markers play a crucial role. Further details of development of chunking system, we have used could be found in (Das and Bandyopadhyay 2009).

4.2.2 Dependency Parser

Dependency depth feature is very useful to identify Theme expressions. A particular Theme word generally occurs within a particular range of depth in a dependency tree. Theme expressions may be a Named Entity, a common noun or words of other POS categories. It has been observed that depending upon the nature of Theme expressions it can occur within a certain depth in the dependency tree in the sentences. A statistical dependency parser has been used for Bengali as described in (Ghosh et al., 2009).

4.3 Discourse Level Features

4.3.1 Positional Aspect

Depending upon the position of the thematic clue, every document is divided into a number of zones. A detailed study was done on the Bengali news

corpus to identify the roles of the positional aspect features of a document (first paragraph, last two sentences) in the detection of theme words. The importance of these positional features has been described in the following sections.

4.3.2 Title Words

It has been observed that the Title words of a document always carry some meaningful thematic information. The title word feature has been used as a binary feature during CRF based machine learning.

4.3.3 First Paragraph Words

People usually give a brief idea of their beliefs and speculations about any related topic or theme in the first paragraph of the document and subsequently elaborate or support their ideas with relevant reasoning or factual information. Hence first paragraph words are informative in the detection of Thematic Expressions.

4.3.4 Words From Last Two Sentences

It is a general practice of writing style that every document concludes with a summary of the overall story expressed in the document. We found that it is very obvious that every document ended with dense theme/topic word in last two sentences.

4.3.5 Term Distribution Model

An alternative to the classical TF-IDF weighting mechanism of standard IR has been proposed as a model for the distribution of a word. The model characterizes and captures the informative-ness of a word by measuring how regularly the word is distributed in a document. Thus the objective is to estimate that measures the distribution pattern of the *k* occurrences of the word w_i in a document d. Zipf's law describes distribution patterns of words in an entire corpus. In contrast, term distribution models capture regularities of word occurrence in subunits of a corpus (e.g., documents, paragraphs or chapters of a book). A good understanding of the distribution patterns is useful to assess the likelihood of occurrences of a theme word in some specific positions (e.g., first paragraph or last two sentences) of a unit of text. Most term distribution models try to characterize the informative-ness of a word identified by inverse document frequency (IDF). In the present work, the distribution pattern of a word within a document formalizes the notion of theme inference informative-ness. This is based on the Poisson distribution. Significant Theme words are identified using TF, Positional and Distribution factor. The distribution function for each theme word in a document is evaluated as follows:

$$f_d(w_i) = \sum_{i=1}^{n} (S_i - S_{i-1}) / n + \sum_{i=1}^{n} (TW_i - TW_{i-1}) / n$$

where n=number of sentences in a document with a particular theme word Si=sentence id of the current sentence containing the theme word and Si-1=sentence id of the previous sentence containing the query term, TW_i is the positional id of current Theme word and TW_{i-1} is the positional id of the previous Theme word.

4.3.6 Collocation

Collocation with other thematic word/expression is undoubtedly an important clue for identification of theme sequence patterns in a document. As we used chunk level collocation to capture thematic words and in this section we are introducing collocation feature as inter-chunk collocation or discourse level collocation with various granularity as sentence level, paragraph level or discourse level.

5 Theme Clustering

Theme clustering algorithms partition a set of documents into finite number of topic based groups or clusters in terms of theme words/expressions. The task of document clustering is to create a reasonable set of clusters for a given set of documents. A reasonable cluster is defined as the one that maximizes the within-cluster document similarity and minimizes between-cluster similarities. There are two principal motivations for the use of this technique in theme clustering setting: efficiency, and the cluster hypothesis.

The cluster hypothesis (Jardine and van Rijsbergen, 1971) takes this argument a step further by asserting that retrieval from a clustered collection will not only be more efficient, but will in fact improve retrieval performance in terms of recall and precision. The basic notion behind this hypothesis is that by separating documents according to topic, relevant documents will be found together in the

same cluster, and non-relevant documents will be avoided since they will reside in clusters that are not used for retrieval. Despite the plausibility of this hypothesis, there is only mixed experimental support for it. Results vary considerably based on the clustering algorithm and document collection in use (Willett, 1988). We employ the clustering hypothesis to measure inter-document level thematic affinity inference on semantics.

Application of the clustering technique to the three sample documents results in the following theme-by-document matrix, A, where the columns represent Docl, Doc7 and Doc13 and the themes politics, sport, and travel.

$$A = \begin{bmatrix} election & cricket & hotel \\ parliament & sachin & vacation \\ governor & soccer & tourist \end{bmatrix}$$

The similarity between vectors is calculated by assigning numerical weights to these words and then using the cosine similarity measure as specified in the following equation.

$$s\left(\vec{q}_k, \vec{d}_j\right) = \vec{q}_k . \vec{d}_j = \sum_{i=1}^{N} w_{i,k} \times w_{i,j} \quad \text{----(1)}$$

This equation specifies what is known as the dot product between vectors. Now, in general, the dot product between two vectors is not particularly useful as a similarity metric, since it is too sensitive to the absolute magnitudes of the various dimensions. However, the dot product between vectors that have been length normalized has a useful and intuitive interpretation: it computes the cosine of the angle between the two vectors.

ID	Theme	1	2	3
1	**প্রশাসন (administration)**	**0.63**	0.12	0.04
1	সুশাসন (good government)	0.58	0.11	0.06
1	সমাজ (society)	0.58	0.12	0.03
1	আইন (law)	0.55	0.14	0.08
2	গবেষণা (research)	0.11	0.59	0.02
2	কলেজ (college)	0.15	0.55	0.01
2	**উচ্চশিক্ষা (higher study)**	0.12	**0.66**	0.01
3	জেহাদি (jehadi)	0.13	0.05	0.58
3	মসজিদ (mosque)	0.05	0.01	0.86
3	নয়াদিল্লী (New Delhi)	0.12	0.04	0.65
3	**কাশ্মীর (Kashmir)**	0.03	0.01	**0.93**

Table 4: Five cluster centroids (mean $\vec{\mu}_j$)

When two documents are identical they will receive a cosine of one; when they are orthogonal (share no common terms) they will receive a cosine of zero. Note that if for some reason the vectors are not stored in a normalized form, then the normalization can be incorporated directly into the similarity measure as follows.

Of course, in situations where the document collection is relatively static, it makes sense to normalize the document vectors once and store them, rather than include the normalization in the similarity metric.

$$s\left(\vec{q}_k, \vec{d}_j\right) = \frac{\sum_{i=1}^{N} w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,k}^2}} \quad \text{----(2)}$$

Calculating the similarity measure and using a predefined threshold value, documents are classified using standard bottom-up soft clustering *k-means* technique. The predefined threshold value is experimentally set as 0.5 as shown in Table 4.

A set of initial cluster centers is necessary in the beginning. Each document is assigned to the cluster whose center is closest to the document. After all documents have been assigned, the center of each cluster is recomputed as the centroid or mean $\vec{\mu}$ (where $\vec{\mu}$ is the clustering coefficient) of its members that is $\vec{\mu} = \left(1/\left|c_j\right|\right) \sum_{x \in c_j} \vec{x}$. The distance function is the cosine vector similarity function.

Table 4 gives an example of theme centroids from the K-means clustering. Bold words in Theme column are cluster centers. Cluster centers are assigned by maximum clustering coefficient. For each theme word, the clusters from Table 4 is still the dominating cluster. For example, "প্রশাসন" has a higher membership probability in cluster 1 than in other clusters. But each theme word also has some non-zero membership in all other clusters. This is useful for assessing the strength of semantic affinity association between a theme word and a topic. Comparing two members of the cluster2, "কাশ্মীর" and "নয়াদিল্লী", it is seen that "নয়াদিল্লী" is strongly associated with cluster2 (p=0.65) but has some affinity with other clusters as well (e.g., p =0.12 with the cluster1). This is a good example of the utility of soft clustering. These non-zero values are still useful for calculat-

ing vertex weight during Semantic Relational Graph generation.

6 Semantic Relational Graph

Representation of input text document(s) in the form of graph is the key to our design principle. The idea is to build a document graph $G=<V,E>$ from a given source document $d \in D$. First, the input document d is parsed and split into a number of text fragments as themes. At this preprocessing stage, text is tokenized, stop words are eliminated, and words are stemmed. Thus, the text in each document is split into fragments and each fragment is represented with a vector of constituent theme words. These text fragments become the nodes V in the document graph.

The similarity between two nodes is expressed as the weight of each edge E of the document graph. A weighted edge is added to the document graph between two nodes if they either correspond to adjacent text fragments in the text or are semantically related by theme words. The weight of an edge denotes the degree of the semantic inference relationship. The weighted edges not only denote document level similarity between nodes but also inter document level similarity between nodes. Thus to build a document graph G, only the edges with edge weight greater than some predefined threshold value are added to G, which basically constitute edges E of the graph G.

The Cosine similarity measure has been used here. In cosine similarity, each document d is denoted by the vector $\vec{V}(d)$ derived from d, with each component in the vector for each Theme words. The cosine similarity between two documents (nodes) *d1* and *d2* is computed using their vector representations $\vec{V}(d1)$ and $\vec{V}(d2)$ as equation (1) and (2) (Described in Section 5). Only a slight change has been done i.e. the dot product of two vectors $\vec{V}(d1) \bullet \vec{V}(d2)$ is defined as $\sum_{i=1}^{M} V(d1)V(d2)$. The Euclidean length of d is defined to be $\sqrt{\sum_{i=1}^{M} \vec{V}_i^2(d)}$ where M is the total number of documents in the corpus. Theme nodes within a cluster are connected by vertex, weight is calculated by clustering co-efficient for those theme nodes and inter cluster vertexes. Cluster centers are interconnected with weighted vertex. The weight is calculated by cluster distance as measured by cosine similarity measure as discussed earlier.

To better aid our understanding of the automatically determined category relationships we visualized this network using the Fruchterman-Reingold force directed graph layout algorithm (Fruchterman and Reingold, 1991) and the NodeXL network analysis tool (Smith et al., 2009)4. A theme relational model graph drawn by NoddeXL is shown in Figure 1.

7 Semantic Distance Measurement

Finally generated semantic relational graph is the desired SemanticNet that we proposed. Generated Bengali SemanticNet consist of almost 90K high frequent Bengali lexicons. Only four categories of POS (noun, adjective, adverb and verb) considered for present generation as reported earlier. In the generated Bengali SemanticNet all the lexicons are connected with weighted vertex either directly or indirectly. Semantic lexicon inference could be identified by network distance of any two nodes by calculating the distance in terms of weighted vertex. We computed the relevance of semantic lexicon nodes by summing up the edge scores of those edges connecting the node with other nodes in the same cluster. As cluster centers are also interconnected with weighted vertex so inter-cluster relations could be also calculated in terms of weighted network distance between two nodes within two separate clusters. As an example:

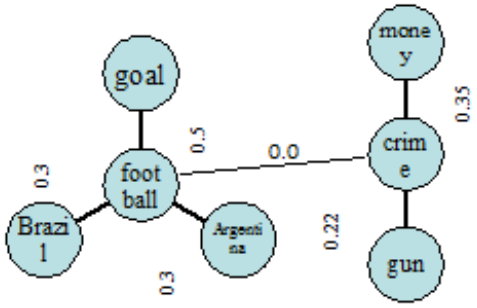

Figure 2: Semantic Affinity Graph

The lexicon semantic affinity inference from Figure 2 could be calculated as follows:

[4] Available from http://www.codeplex.com/NodeXL

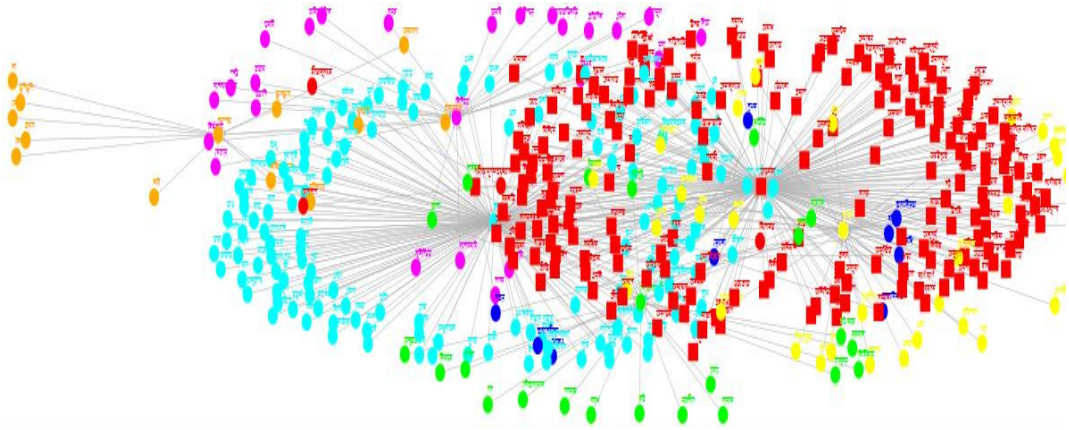

Figure 1: Semantic Relational Graph by NodeXL

$$S_d(w_i, w_j) = \frac{\sum_{k=0}^{n} v_k}{k} \quad \text{----(1)} \quad \text{or}$$

$$\sum_{c=0}^{m} \frac{\sum_{k=0}^{n} v_k}{k} \times \prod_{c=0}^{m} l_c \text{---(2)}$$

where $S_d(w_i, w_j) =$ semantic affinity distance between two lexicons w_i and w_j. Equation (1) and (2) are for intra-cluster and inter-cluster semantic distance measure respectively. *k*=number of weighted vertex between two lexicons w_i and wj. v_k is the weighted vertex between two lexicons. *m*=number of cluster centers between two lexicons. l*c* is the distance between cluster centers between two lexicons.

For illustration of present technique let take an example:

$$(\text{Argentina, goal}) = \frac{0.5 + 0.3}{2} = 0.4$$

$$(\text{Gun, goal}) = \left(\frac{0.22}{1} + \frac{0.5}{1}\right) \times 0.0 = 0$$

,

It is evident from the previous example that the score based semantic distance can better illustrate lexicon affinity between Argentina and goal but is no lexicon affinity relation between gun and goal. Instead of giving only certain semantic relations like WordNet or ConceptNet the present relative probabilistic score based lexicon affinity distance based technique can represent best acceptable solution towards represent human pragmatic knowledge. Not only ideologically rather the SemanticNet provide a good solution to any type of NLP problem. A detail analysis of an Information retrieval system and Summarization using SemanticNet has been detailed in evaluation section.

Although every lexicon pair cannot be labeled by exact semantic role but we try to keep a few semantic roles for a crossroad from previous techniques to this new one. These semantic relations may be treated as a bridge to traverse SemanticNet by gathering knowledge from other resources WordNet and ConceptNet. Approximately 22k (24% of overall SemanticNet) lexicons are tagged with appropriate semantic roles by these two processes described below.

8 Semantic Role Assignment

Two types of methods have been taken to assign pair wise lexicon semantic affinity relations. First one is derived from ConceptNet. In the second technique sub-graph is identified consisting of a nearest verb and roles are assigned accordingly.

8.1 Semantic Roles from ConceptNet

A ConceptNet API[5] written in Java has been used to extract pair wise relation from ConceptNet. A Bengali-English dictionary (approximately 102119 entries) has been developed using the Samsad Bengali-English dictionary[6] used here for equivalent lookup of English meaning of each Bengali lexicon. Obtained semantic relations from ConceptNet for any lexicon English pair are assigned to source Bengali pair lexicons. As an example: (“Tree”,”Gree”) (“গাছ”,”সবুজ”): *OftenNear, Partof, PropertyOf, IsA*.

[5] http://web.media.mit.edu/~hugo/conceptnet/

[6] http://dsal.uchicago.edu/dictionaries/biswas_bengali/

8.2 Verb Sub-Graph Identification

It is an automatic process using manually augmented list of only 220 Bengali verbs. This process starts from any arbitrary node of any cluster and start finding any nearest verb within the cluster. As we reported earlier there is no POS information in the generated semantic relational graph so system uses the manually augmented list of verbs as partly reported in Table 5.

Verb	English Gloss	Probable Relations
হয়	Be	IsA
আছে	Have	CapableOf
থাকা	Have	CapableOf
তৈরি	Made	MadeOf
বসবাস	Live	LocationOf

Table 5: Relations

The relation label attached with every verb in the manually augmented list (as reported in Table 5) is then automatically assigned between each pair of lexicons.

9 Evaluation

It is bit difficult to evaluate this type of lexicon resources automatically. Manual validation may be suggested as a better alternative but we prefer for a practical implementation evaluation strategy.

9.1 Information Retrieval (IR)

For evaluation of Bengali SemanticNet it is used in Information Retrieval task using corpus from Forum for Information Retrieval Evaluation (FIRE)[7] ad-hoc mono-lingual information retrieval task for Bengali language. Two different strategies have been taken. First a standard IR technique with TF-IDF, zonal indexing and ranking based technique (Bandyopadhyay et al., 2008) has been taken. Second technique uses more or less same strategy along with query expansion technique using SemanticNet (Although the term SemanticNet was not mentioned there) as a external lexicon resource (Bhaskar et al., 2010).

Only the following evaluation metrics have been listed for each run: mean average precision (MAP), Geometric Mean Average Precision (GM-AP), (document retrieved relevant for the topic) R-Precision (R-Prec), Binary preferences (Bpref) and Reciprical rank of top relevant document (Recip_Rank). The evaluation strategy follows the global standard as Text Retrieval Conference (TREC)[8] metrics. It is clearly evident from the system results reported in Table 6 that SemanticNet is a better way to solve lexicon semantic affinity.

Scores	Bengali IR using	
	IR	Using SemanticNet
MAP	0.0200	0.4002
GM_AP	0.0004	0.3185
R-Prec	0.0415	0.3894
Bpref	0.0583	0.3424
Recip_Rank	0.4432	0.6912

Table 6: Information Retrieval using SemanticNet

Evaluation result shows effectiveness of developed SemanticNet in IR. Further analysis revealed that general query expansion technique generally used WordNet synonyms as a resource. But in reality "হৃদয়" and "পরাণ" could not be clustered in one cluster though they represent same semantic of 'heart'. First one used in general context whereas the second one used only in literature. If there is any problem to understand Bengali let come with an example of English. Conceptually "you" and "thy" could be mapped in same cluster as they both represent the semantic of 2nd person but in reality "thy" simply refers to the literature of great English poet Shakespeare. Standard lexicons cannot discriminate this type of fine-grained semantic differences.

9.2 Summarization

To measure the effectiveness of SemanticNet into multiple areas in NLP we tried with automatic summarization technique. A gold standard Bengali corpus of 100 NEWS documents (collected from www.anandabazar.com) has been created. Three annotators (Mr. X, Mr. Y and Mr. Z) participated in the present task. Annotators are asked to extract those sentences from documents that best describe the concise thematic notion of any document. The agreement of annotations among three annotators has been evaluated as reported in Table 7.

Annotators	**X vs. Y**	**X Vs. Z**	**Y Vs. Z**	**Avg**
Percentage	73.87%	69.06%	60.44%	67.8%
All Agree	58.66%			

Table 7: Agreement of annotators at sentence level

With this data we tried a standard Bengali summarization technique as described in (Paladhi and

[7] http://www.isical.ac.in/~clia/index.html

[8] http://trec.nist.gov/

Bandyopadhyay) (1) and later on this method used SemanticNet (2) as an external lexical resource. The result of the both system is reported in Table 8.

	Metrics		X	Y	Z	Avg
Summarization	**Precision**	1	72.44%	62.81%	64.55%	66.60%
		2	77.65%	67.22%	71.57%	**72.15%**
	Recall	1	60.03%	57.68%	54.25%	57.32%
		2	68.76%	64.53%	68.68%	**67.32%**
	F-Score	1	66.6%	59.87%	58.36%	61.61%
		2	72.94%	65.85%	70.10%	**69.65%**

Table 8: Summarization with and without SemanticNet

10 Conclusion and Future Task

Experimental result of Information Retrieval and Automatic Summarization using SemanticNet proves it is the best solution rather than any existing lexicon resources. The development strategy employs less human interruption rather a general architecture of Theme identification or Theme Clustering technique using easily extractable linguistics knowledge. The proposed technique could be replicated for any new language.

SemanticNet could be useful any kind of Information Retrieval technique, Information Extraction technique, and topic based Summarization and we hope for newly identified NLP sub disciplines such as Stylometry or Authorship detection and plagiarism detection etc.

Our future task will be in the direction of different experiments of NLP as mentioned above to profoundly establish the efficiency of SemanticNet over multiple domains. Furthermore we will try to develop SemanticNet for many other languages.

References

Bandyopadhyay S., Das A., Bhaskar P.. English Bengali Ad-hoc Monolingual Information Retrieval Task Result at FIRE 2008. In Working Note of Forum for FIRE-2008.

Bhaskar P., Das A.,Pakray P.and Bandyopadhyay S.(2010). Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010, In FIRE-2010.

Das A. and Bandyopadhyay S. (2009). Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII).

Das A. and Bandyopadhyay S. (2010). Morphological Stemming Cluster Identification for Bangla, In Knowledge Sharing Event-1: Morphological Analyzers and Generators, Mysore.

Ekbal A., Bandyopadhyay S. (2008). A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation Journal. Pages 173-182, 2008

Fillmore Charles J., Johnson Christopher R., and Petruck Miriam R. L.. 2003. Background to FrameNet. International Journal of Lexicography, 16:235–250.

Fruchterman Thomas M. J. and Reingold Edward M.(1991). Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164.

Ghosh A., Das A., Bhaskar P., Bandyopadhyay S.(2009). Dependency Parser for Bengali: the JU System at ICON 2009. In NLP Tool Contest ICON 2009, December. Hyderabad.

Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. In Proc. Information Storage and Retrieval, 7, 217-240.

Kipper Karin, Korhonen Anna, Ryant Neville, and Palmer Martha. Extending VerbNet with Novel Verb Classes. In LREC 2006. Genoa, Italy. June, 2006.

Liu Hugo and Singh Push (2004). ConceptNet: a practical commonsense reasoning toolkit. BT Technology Journal, 22(4):211-226.

Palmer Martha, Gildea Dan, Kingsbury Paul, The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics Journal, 31:1, 2005.

Paladhi Sibabrata and Bandyopadhyay Sivaji (2008). "A Document Graph Based Query Focused Multi-Document Summarizer". Proceedings of the 2nd International Workshop on Cross Lingual Information Access (CLIA), pp. 55-62

Smith M, Ben S, Natasa MF, Eduarda R, Vladimir B, Cody D, Tony C, Adam P and Eric G. 2009. Analyzing (social media) networks with NodeXL. LNCS. Springer.

Willerr, P. (1988). Recent trends in hierarchic document clustering: A critical review. Information Processing and Management, 24(5), 577-597.

F Vargha-Khadem, K Watkins, K Alcock, P Fletcher, and R Passingham. Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder, In Proc Nat Acad Sci USA 92, 930 – 933 (1995).