International Journal of Computational Linguistics and Applications vol. 7, no. 2, 2016, pp. 65–85
Received 08/02/2016, accepted 09/03/2016, final 19/09/2016
ISSN 0976-0962, http://ijcla.bahripublications.com

Measuring Semantic Textual Similarity of Sentences Using Modified Information Content and Lexical Taxonomy

GOUTAM MAJUMDER[1], PARTHA PAKRAY[1],
AND ALEXANDER GELBUKH[2]

[1] *National Instutute of Technology Mizoram, India*
[2] *Instituto Politécnico Nacional, Mexico*

ABSTRACT

In this paper, we present a survey and comparative studies on semantic textual similarity methods, those are based on WordNet taxonomy. We also proposed a new method for measuring semantic similarity between sentences. This proposed method, uses the advantages of taxonomy methods and merge these information to a language model. It considers the WordNet synsets for lexical relationships between nodes/words and uni-gram language model is implemented over a large corpus to assign the information content value between the two nodes of different classes. Finally, a similarity score is generated by considering the maximum weight and shortest distance of the graph. To evaluate and compare the method, SemEval 2015 English STS task 2 training dataset is considered.

KEYWORDS: *WordNet Taxonomy, Natural Language Processing, Semantic Textual Similarity, Information Content, Random Walk*

1 INTRODUCTION

In Natural Language Processing (NLP), semantic similarity plays an important role and one of the fundamental tasks for many NLP applications and its related areas. During the evolution of semantic textual similarity (STS), which is defined by a metric over a set of documents where the idea is to find the similarity between them. Similarity between the documents is based on direct and indirect relationships among them [1], [2]. These relationships can be measured and recognized by the presence of semantic relations among them. Identification of STS in short text was proposed in 2006 with the works reported in [3], [4]. After that, focus was shifted on large documents or individual words.

After that, since 2012 the task of semantic similarity is not only limited to find out the similarity between two texts, but also to generate a similarity score from 0 to 5 by different SemEval tasks[3] [5–7]. In this task, a scale of 0 means unrelated and 5 means complete semantically equivalence.

Since its inception, the problem was seen a large number of solutions in a relatively small amount of time. The central idea behind the most solution is that, the identification and alignment of semantically similar or related words across the two sentences and the aggregation of these similarities to generate an overall similarity [8, 4, 9].

One of the major goals of STS task, is to create a unified framework by combining several independent semantic components to find their impact over several NLP tasks. Developing such framework is an important research problem having many important applications in NLP such as information retrieval (IR) and in digital education like text summarization [10, 11], question answering [12], relevance feedback and text classification [13], word sense disambiguation [3], and extractive summarization [14].

Semantic similarity also contributes for many Semantic Web applications like community extraction, ontology generation and entity disambiguation. It is also useful for Twitter search [14], where requires the ability to accurately measure semantic relatedness between concepts or entities. In IR one of the main problems is to retrieve a set of documents and retrieving images by captions [15],

[3] http://ixa2.si.ehu.es/stswiki/index.php/Main_Page

which is semantically related to a given user query in a web search engine.

In database also, text similarity can be used for schema matching to solve the semantic heterogeneity for data sharing system, data integration system, message passing system, and peer-to-peer data management system [16]. It is also useful for relational join operations in database where join attributes are textually similar to each other. It has a verity of application domain including integration and querying of data from heterogeneous resources, cleaning of data, and mining of data [17, 18].

In NLP it is also noticed that, STS is related to both Textual Entailment (TE) and Paraphrasing, but differs in number of ways. In NLP, TE can draws three directional relationships between two text fragments while the task considered two text fragments as text (t) and hypothesis (h) respectively. On the other hand paraphrasing identification is the task of recognizing text fragments with approximately the same meaning within a specific context. So TE and paraphrasing gives a yes/no decision and STS identifies the degree of equivalence of text and rated them on the basis of their semantic relationships.

Measuring semantic similarity between texts can be categories in following ways, such as (i) topological (ii) statistical similarity (iii) semantic based (iv) vector space model (v) word alignment based and (vi) machine learning. Among these methods, topological studies plays an important role to understand intended meaning of an ambiguous word, which is very difficult to process computationally. For many NLP related task it is important to understand the semantic relation between the word/concepts. To decompose such systems we need to work with word level relation and those can be considered as hierarchical, associative and equivalence.

In this work, we analyse the different topological based methods, those were already proposed for identification of words class such as *noun* using WordNet synsets. We also proposed a new method for detection of textual similarity between sentences based on language model and WordNet taxonomy. The complete illustration of the proposed method is shown in section 4. The literature review on taxonomy is reported in second section and in the third section, a complete illustration of different taxonomy methods is presented. In the sixth section results of a short experiment is reported and compared. Finally conclusion of the work is reported in the last section.

2 Literature Review

In many cases determining the intended meaning of an ambiguous word is difficult for human and it is quite difficult to process automatically also. This ambiguity can be eliminated by considering following relationships among the words or concepts: (i) hierarchical (e.g. IS-A or hypernym-hyponym, part-whole etc.), (ii) associative (e.g. cause-effect) and, (iii) equivalence [19]. Among these IS-A relation is widely used and studied, which maps to the human cognitive view of classification (i.e. taxonomy). The IS-A relation among the concepts has been suggested and employed as a special case of semantic similarity of distance [20]. Semantic similarity can be estimated by defining a topological similarity by using ontologies to define the distance between term and concepts.

Taxonomy is often represented as a hierarchical structure and also considered as a network structure. To measuring the similarity information of this network can be useful. There are several ways to determine the conceptual similarity between two words in a hierarchical semantic network. There are essentially two types of approaches which, calculate topological similarity between ontological concepts. Those are (i) node based (information-content approach) and (ii) edge-based (distance based).

Issues related to lexical association was reported in [21], where a generalization technique of lexical association was proposed. To solve these issues (i.e. reliable word/word correspondence) author facilitate different statistical facts by considering word classes rather than individual words. In this task a set of possible word classes were constructed from WordNet [22] and an investigation was conducted to identify the relationship between word/classes using mutual information. For word-based information retrieval information from WordNet was passed over a SMART environment where a content description was added (only part-of-speech information) with the input text.

Ambiguity of word form during document indexing was investigated using a semantic based network where semantic distance between network nodes was considered [23]. In this work, word sense during document indexing was studied using the WordNet semantic network. Distance between multiple senses of input word was disambiguated by finding the combination of senses from a set of contiguous terms.

In another work, Rensik. P proposed a method for identification of semantic similarity in a taxonomy based on the notion of information content [21]. Similarity between two words/concepts was evaluated by considering the common information between them and a set of fifty thousand (50,000) node form WordNet taxonomy of noun class was considered for this task. To calculate the frequencies of concepts Brown Corpus of American English (having 1000, 000 words) was considered [24].

Jiang and Conrath introduces a new approach for measuring semantic similarity between words using lexical taxonomy structure with corpus statistical information. So the semantic distance between nodes in the semantic space was constructed by the taxonomy, which provides a better result with the computational evidence those are derived from a distributional analysis of corpus data. This proposed method, is a combined approach in which edge counting scheme was inherited and further enhanced by node based approach [19].

To find the similarity between phrases and sentences a random walk over a graph was proposed [25]. In this work, local semantic information and semantic resources of WordNet was combined together. Semantic signature generated by random walk was compared to another such distribution to get the similarity score. They also showed that, graph work similarity between texts is a feature for recognizing textual entailment also.

Methods introduced in this section are based on topological similarity between ontological concepts and apart from these methods related to ontological instances namely: (i) pair-wise; and (ii) group-wise also found. It was founded that methods based on ontological instances are mainly used to represent medical knowledge and no such work was noticed, which was used for semantic textual similarity between classes or phrases or sentences. So all these task are not reported here, because proposed work is planned for textual similarity only. In the next section a detailed illustration is presented for methods used to identify the semantic relatedness between words/classes based on taxonomical concepts.

3 Methodology

In the field of Information Retrieval (IR), document retrieval based on semantic similarity of words has been largely investigated and, all these methods consider the semantic and ontological relationships that exist between the words (e.g. polysemy, synonym etc.). So based on this knowledge semantic similarity between objects in ontology can broadly categorise into three groups like: (i) node-based; (ii) edged-based; and (iii) hybrid where it combines node and edge-based.

3.1 *Node Based (Information Content) Approach*

Node based approaches also called Information Content (IC) approaches [26, 21], which is used to determine the semantic similarity between concepts. In this method, each of the concept or node poses IS-A taxonomy are kept in one set called C and all of these nodes carry unique concepts. Intuitively, one key to the similarity of two concepts is that to which they share information in common. In taxonomy direct relation between two concepts can be found by an edge counting method. In this method, if the minimal path between two nodes is long, that means it is necessary to go high in the hierarchy to find a least upper bound. An example of IS-A relationship between the concepts is shown in Fig. 1, where two concepts *NICKEL* and *DIME* both subsumes *COIN*. In this example *NICKEL* and *CREDIT CARD* shares a common super class *MEDIUM OF EXCHANGE* [27].

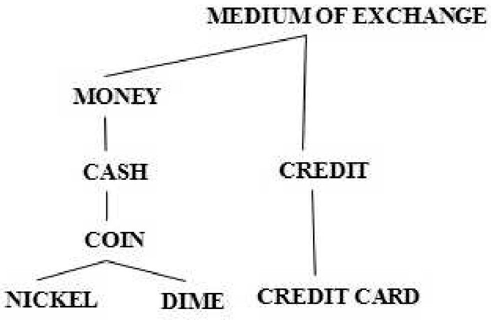

Fig. 1. Representation of a WordNet Taxonomy (IS-A Relationship)

To avoid unreliability edge-distances between nodes, it is possible to associates probability with taxonomy. The value of IC of a class is obtained by estimating the probability in a large corpus with a function $p : C \rightarrow [0, 1]$ if $c \in C$, $p(c)$, is the probability of encountering an instance c. Considering the notation of information theory [28], IC of a class can be calculated as follows:

$$IC(c) = log^{-1}p(c) \tag{1}$$

Quantifying information content in this way: if the probability increases, its information content decreases. It means that if there is a unique top in the tree, then its probability is 1 so the information content is 0 and the similarity of two concepts can be calculated as follows:

$$sim\left(c_1, c_2\right) = max_{c \in S(c_1, c_2)}\left[-log\ p\left(c\right)\right] \tag{2}$$

where $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2. From Fig. 1, it is identified that, similarity of *NICKEL* and *DIME* can be calculated by considering all the upper bounds. Among those upper bounds node having highest information content value is considered as similarity score between these two nodes.

To implement the information content model reported in [11], WordNet fifty thousand nodes were considered, where taxonomy of concepts represented by nouns and compound nominals [22]. Before implementing IC, two concepts need to define as sets of *words(c)* and $classes(w)$. $Words(c)$ is the set of words subsumed by the class and $classes(w)$ is defined as the classes in which the word is contained. The class can be seen as a sub-tree in the whole hierarchy and $classes(w)$ is the set of possible senses that the word has:

$$classes\,(w) = \{c|w \in words\,(c)\} \tag{3}$$

A simple class/concept frequency formula is also defined in [21] and [29], where the number of word sense factor:

$$freq\,(c) = \sum_{w \in words(c)} freq\,(w) \tag{4}$$

and

$$freq\,(c) = \sum_{w \in words(c)} \frac{freq\,(w)}{|classes\,(c)|} \tag{5}$$

Finally, the class probability can be computed using maximum likelihood estimation (MLE):

$$p\,(c) = \frac{freq\,(c)}{N} \tag{6}$$

where N is the total number of nouns observed (excluding those not subsumed by any WordNet class, of course).

An example of multiple inheritance concepts like *NICKLE* and *GOLD* those have more super classes as shown in Fig. 2. In this case one word have more sense, so the similarity can be determined by the best similarity value among all the class pairs, which belongs to their various senses [19]:

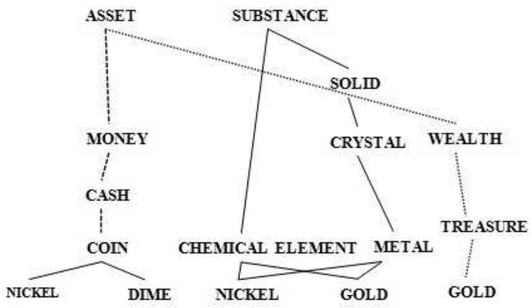

Fig. 2. WordNet Taxonomy of Multiple Inheritance

$$sim\left(w_1, w_2\right) = max_{c_1 \in sen(w_1), c_2 \in sen(w_2)} \left[sim\left(c_1, c_2\right)\right] \qquad (7)$$

where $sen(w)$denotes the set of possible senses for word w.

In another task [25], extends node based method to vector space model for semantic measure using random walk algorithm. In this approach, instead of comparing between two text segments directly, it compares distribution of each text and a random walk is generated over a graph, which is derived from WordNet and corpus statistics.

WordNet is itself a graph over clusters, which contains one sense of one or more similar words. Each node in the graph represents a synset. Word having different meaning: multiple synsets (or cluster) is generated based on different meaning. For example the word *bank* belongs to the two different synsets, one for financial bank and other for river bank. By constructing each edge created from a WordNet relationship is guaranteed to have a corresponding edge in the opposite direction. Nodes are connected with edges (represents the relation) corresponding too many relationships from WordNet is as follows: hypernym/ hyponym, instance/ instanceof, all holonym/meronym links, antonym, entails/entailed by, adjective satellite, causes/caused by, participle, pertains to, derives/derived from, attribute/has attribute, and topical (but not regional or usage) domain links. Following types of nodes from WordNet was considered for graph construction:

- **Synset:** Each WordNet synset has a corresponding node. For example, one node corresponds to the synset referred to by "dog#n#3," the third sense of dog as noun, whose meaning is "an informal term for a man."
- **TokenPOS:** One node is allocated to every word coupled with a part of speech, such as "dog#n" meaning dog as a noun. These nodes link to all the synsets they participate in, so that "dog#n" links the synset nodes for canine, hound, hot dog etc.
- **Token:** Nodes do not have any part-of-speech information in synsets, all the *TokenPOS* nodes were linked with all these tokens.

Random walk methods have following advantages over traditional node based method:

- It enables the similarity measure to have a principled means by combining multiple types of edges from WordNet.
- By traversing all the links, random walk aggregates the local similarity statistics across the entire graph [30].

A random walk of an undirected weighted graph was defined with transition probability between the links of the elements of database, which is designed with WordNet. So, a random walker can jump from element to element and each element of Markov-chain represents a state into the taxonomy. Finally, the similarities between text passages had been computed using Markov-chain Model [31].

In Markov-chain model, a weighted graph G with weight w_{ij} between node i and j was considered, where the database elements and links represents node and edges of the graph. The weight w_{ij} must have following convention: the relation between i and j is more, the larger the value of w_{ij} and the walk should be minimum and the value of $w_{ij} > 0$ and $w_{ij} = wji$.

Sequence of node visited by a random walker are called a random walk and described by a Markov-chain. A random variable $s(t)$ contains the current state of the Markov chain at time t: if the random walker is in state i at time t, then $s(t) = i$. The random walk is defined with the following single-step transition probabilities of jumping from any state or node $i = s(t)$ to an adjacent node j as follows:

$$j = s\left(t+1\right) : P\left(s\left(t+1\right) = j | s\left(t\right) = i\right) = \frac{a_{ij}}{a_{i.}} = P_{ij} \tag{8}$$

where $a_{i.} = \sum_{j=1}^{n} a_{ij}$ and a_{ij} is the the elements of symmetric adjacency matrix A of the graph and defined as $a_{ij} = w_{ij}$, if i and j is connected else 0.

We need to compare the stationary distribution of two Markov chains of two text passages to calculate the semantic relations between them. The transition probability $n_i^{(t)}$ of finding the particle of any node as the sum of all ways in which the particle could have reached n_i from any other node at the previous time step as follows:

$$n_i^{(t)} = \sum_{n_i \in V} {n_j}^{(t-1)} P\left(n_i | n_j\right) \tag{9}$$

where $P(n_i|n_j)$ is the probability of transitioning from n_j to n_i.

3.2 *Edge-based (Distance) Approach*

The edge based approach is the direct way of computing semantic similarity in taxonomy. It counts the number of edges between two nodes those corresponds to the concepts being compared. Minimum the path between two nodes they are more similar [19].

It was pointed out that, in a hierarchical taxonomy distance between nodes must satisfy the matrix properties like: zero property, semantic property, positive property. Because of distance between two adjacency nodes should not necessary equal, so it is necessary an edge connecting two nodes must be weighted. To determine the weight following structural characteristics should be considered [20]:

- **Network Density:** higher densities in WordNet need to consider for example plant-flora section in WordNet for measuring the network density. Distance between the nodes is closer to the local density which is reported in [29].
- **Node Depth:** in terms of the depth it can be said that distance shrinks as one descends down a hierarchy.
- **Type of link:** it represents the relation between two nodes. In many edge-based model only IS-A link is consider [32, 20]. Other relations can also consider such as Meronym-Holonym, which have different effect for calculating the weight.
- **Link Strength of specific child link:** this could be measured using WordNet relationships between child node and its parent node.

Weight measurement also done manually for the edges those are reported in [33, 32, 20, 34]. To measure weight automatically, certain observations were considered over the Hierarchical Concept Graph (HCG). For measuring the weight of a link, density, depth of the HCG and link strength between child and parent nodes is considered [29]. The density of a HCG for a specific link type is estimated by counting the number of links of that type. The strength between the links was estimated as a function of nodes IC value and its sibling and parents node. Finally, results of these two operations were normalized by dividing the depth of the link.

A minimum and maximum range was taken before measuring the weight between two nodes [23]. Because of an edge represents two inverse relations the final weight of an edge was fixed by averaging the two weight values. The depth-relative scaling process was adopted in which the average value is divided by the depth of the edge within the overall tree. The weight of an edge of two adjacent nodes n_1 and n_2 was calculated following ways:

$$w\left(n_1, n_2\right) = \frac{w\left(n_1 \rightarrow_r n_2\right) + w\left(n_2 \rightarrow_{r'} n_1\right)}{2d} \tag{10}$$

given

$$w\left(X \rightarrow_r Y\right) = max_r - \frac{max_r - min_r}{n_r\left(X\right)} \tag{11}$$

where $\rightarrow_r$ is a relation of type r, and $\rightarrow_{r'}$ is its reverse, d is the depth of the deeper one of the two and $n_r\left(X\right)$ is the number of relations of type r leaving node X.

The value of 0 was assigned for all synonym type relations. Holonymy, hyponymy, hypernymy, and meronymy are the types of relation, where weights ranging from 1 to 2 and for antonymy type relation weights was assigned as 2.5.

Edge counting method has been considered to determine the edge based similarity [21]. To convert the distance measure to the similarity measure, by subtracting the path length from the maximum possible path length as follows:

$$sim_{edge}\left(w_1, w_2\right) = 2d_{max} - \left[min\left(c_1, c_2\right) len\left(c_1, c_2\right)\right] \tag{12}$$

where d_{max} represents the maximum depth in the taxonomy, and then c_1 and c_2 ranges over senses of word w_1 and w_2 respectively.

3.3 *Hybrid Approach*

Node and edge based methods discussed in previous sections have many differences in between them. The edge-based methods, looks true without any concise reasoning and on the other hand, node-based approach looks more accurate than distance-based. The distance measure was relayed on the subjective knowledge of the network while the WordNet was used not for measuring the similarity, but for construction of the network layers.

On the other side information content was not sensitive to the link types [21], but it is dependent on the structure of the taxonomy. Although these two methods are different from each other, a combined method was derived from edge-based while it considers the information content as a decision factor [19].

In this method, link strength factor was first considered by taking the conditional probability of the child concept c_i of its parent concepts p:

$$P\left(c_i|p\right) = \frac{c_i \cap p}{P\left(p\right)} = \frac{P\left(c_i\right)}{P\left(p\right)} \tag{13}$$

The link strength (LS) was defined by considering the negative logarithm of the conditional probability (see equation 4), by following the argument of information theory (see equation 1) as follows:

$$LS\left(c_i, p\right) = -log\left(P\left(c_i|p\right)\right) = IC\left(c_i\right) - IC\left(p\right) \tag{14}$$

Form (equation 14) it is clearly understood that the difference of information content values between child and parent has been considered as *LS*.

By considering other structural characteristics mentioned edge-based approach also considered here to calculate the weight wt of a child node as follows:

$$wt\left(c, p\right) = \left(\beta + \left(1 - \beta\right)\frac{\bar{E}}{E\left(P\right)}\right)\left(\frac{d\left(p\right) + 1}{d\left(P\right)}\right)^{\alpha}\left[IC\left(c\right) - IC\left(p\right)\right]T\left(c, p\right) \tag{15}$$

where $d\left(p\right)$,$E\left(P\right)$denotes the depth, local density of the node p

respectively and $\bar{E}$ represents the average density in the tree and $T\left(c,p\right)$ is the link type factor. The parameters α and β controls the degree of depth and density to calculate the edge weight. So the distance between two nodes is the summation of edge weights and a shortest path between them.

$$Dist\left(w_1,w_2\right) = \sum_{c\in\{path(c_1,c_2)-LSuper(c_1,c_2)\}} wt\left(c,parent\left(c\right)\right) \tag{16}$$

4 PROPOSED METHOD

In this paper, a language model based semantic network has been proposed to find the semantic similarity between two English sentences. Among these two sentences one is considered as source as S and other as target text T. We assume that both S and T are syntactically and semantically correct. The proposed system can be brought down into following stages:

- In any language processing it is important to remove all the stop words before start any semantic similarity task. Initially all the stop words, have stored in a Java array and after that all the words of S and T is considered one after another for identification. Although stop words are most commonly used words but there is no universal list available for all language processing task[4]. These identified stop words are ignored during similarity stage.
- In first step, Peen Treebank tag set [35] is used to label the words for part-of-speech (POS) information, which is most commonly used syntactic information. Further these identified tags and words are input to the system to generate the parse tree.
- To generate the parse tree top down parsing is followed by considering its advantages over the bottom up parsing. For parsing

[4] http://xpo6.com/list-of-english-stop-words/

all English grammatical role is considered. After that identified phrase structures is used to generate the top-down parse tree.

- In this stage, a multi-stage (equal to level of the tree) undirected weighted graph is designed by considering the parse tree along with other statistical information found in the previous stages. Following characterises is considered for graph construction:
  - **Part-of-Speech:** All the stop words based on its POS information is not considered, when two words are found same in two parse tree at same level.
  - **Node Depth:** Starting from the root node S all possible paths are considered till the search ends with a word/concept at higher lever (i.e. leaf node) of the tree. The depth of any word is consider in the similarity measuring stage when a word is found in both the parse tree at same level and shares same POS tag.
  - **String Matching:** If any word is found in the parse tree of S and T, which possess *nnp* as POS tag then a weight value to the link is assigned if both the node are same.
- After the completion of graph construction stage weight is measured between the nodes of two graphs. Assigning of weight is performed under the following condition:
  - if POS tag is found different of two nodes of same level then WordNet taxonomy relationship is considered. To calculate the information content i.e. weight w_i of the link the negative logarithm of the conditional probability (see equation 4) as well as argument of information theory is considered.
  - if POS tag is different but strings are matched then two different weight values are calculated.

$$w_i^1 = sim\left(c_1, c_2\right) \tag{17}$$

and

$$w_i^2 = freq_{counts}\frac{c_i}{N} \tag{18}$$

where c_1 and c_2 represents two concepts of two parse tree at same level. N represents as total number of words along with POS tag from a large text corpus and c_i represents total of class c. Finally, the maximum of w_i^1 and w_i^2 is considered for weight.

– if no condition matched and phrase is identified as noun class and words are proper noun then no weight is measured for the link between the current node and proper noun node.
– Finally, similarity is calculated as the minimum distance path while considering maximum weight of the link. After that, an average is calculated by summing of all weights of links starting form start node S till the leaf node.

5 Experimental Results

In order to evaluate the text similarity measure, pair of 50 sentences is taken from SemEval 2015 training dataset[5]. For this task, two different runs are conducted. For the first run, we consider WordNet taxonomy relationships and 0.46 similarity score is reported in this run. For this task WordNet version 2.0 is considered. In second run, we improved the similarity score using information content. For this task highest score is 0.78. In this method, we calculate the IC value by the combining of WordNet taxonomy and uni-gram language model, which out performs the other methods reported in [30], [19] and [26].

6 Conclusion

From this work, it is clearly understood that, node based approaches fully depends on the information content value between two nodes and distance based approaches depends on the depth of semantic network. On the other side, hybrid method works with weight value between child and parent nodes to find the similarity of two classes.

The proposed method, which uses the uni-gram model and hybrid method for measuring the weight between two nodes, which uses the advantages of WordNet information like node-based and distance-based approach. Finally, to measure the similarity the minimum path of the graph and maximum weight of link is considered for generating the similarity score.

[5] http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools

ACKNOWLEDGMENTS This work presented here under the research project Grant No. YSS/2015/000988 and supported by the Department of Science & Technology (DST) and Science and Engineering Research Board (SERB), Govt. of India. Authors are also acknowledges the Department of Computer Science & Engineering of National Institute of Technology Mizoram, India for proving infrastructural facilities.

REFERENCES

1. Corley, C., Mihalcea, R.: Measuring the semantic similarity of texts. In: Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment, Association for Computational Linguistics (June 2005) 13–18
2. Rus, V., Lintean, M.C., Banjade, R., Niraula, N.B., Stefanescu, D.: Semilar: The semantic similarity toolkit. In: ACL (Conference System Demonstrations). (August 2013) 163–168
3. Li, Y., McLean, D., Bandar, Z.A., O'shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE transactions on knowledge and data engineering **18**(8) (2006) 1138–1150
4. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. AAAI **6** (July 2006) 775–780
5. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., Wiebe, J.: Semeval-2014 task 10: Multilingual semantic textual similarity. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). (2014) 81–91
6. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In: In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics, Citeseer (2013)
7. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: Semeval-2012 task 6: A pilot on semantic textual similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. (June 2012) 385–393
8. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD) **2**(2) (2008) 10
9. Šarić, F., Glavaš, G., Karan, M., Šnajder, J., Bašić, B.D.: Takelab: Systems for measuring semantic text similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings

of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. SemEval '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 441–448

10. Aliguliyev, R.M.: A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications **36**(4) (2009) 7764–7772
11. Steinberger, J., Jezek, K.: Using latent semantic analysis in text summarization and summary evaluation. In: Proc. ISIM'04. (April 2004) 93–100
12. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. (June 2011) 752–762
13. Rocchio, J.J.: Relevance feedback in information retrieval. Prentice-Hall, Englewood Cliffs NJ (1971)
14. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Information Processing & Management **33**(2) (1997) 193–207
15. Coelho, T.A., Calado, P.P., Souza, L.V., Ribeiro-Neto, B., Muntz, R.: Image retrieval using multiple evidence ranking. IEEE Transactions on Knowledge and Data Engineering **16**(4) (2004) 408–417
16. Halevy, A.Y., Ives, Z.G., Madhavan, J., Mork, P., Suciu, D., Tatarinov, I.: The piazza peer data management system. IEEE Transactions on Knowledge and Data Engineering **16**(7) (2004) 787–798
17. Cohen, W.W.: Data integration using similarity joins and a word-based information representation language. ACM Transactions on Information Systems (TOIS) **18**(3) (2000) 288–321
18. Schallehn, E., Sattler, K.U., Saake, G.: Efficient similarity-based operations for data integration. Data & Knowledge Engineering, Elsevier **48**(3) (2004) 361–387
19. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: arXiv preprint cmp-lg/9709008. (September 1997)
20. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics **19**(1) (1989) 17–30
21. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
22. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38**(11) (1995) 39–41
23. Sussna, M.: Word sense disambiguation for free-text indexing using a massive semantic network. In: Proceedings of the second international conference on Information and knowledge management, ACM (December 1993) 67–74

24. Kucera, H., Francis, W.N.: Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin (1982)
25. Ramage, D., Rafferty, A.N., Manning, C.D.: Random walks for text semantic similarity. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing, Association for Computational Linguistics (2009) 23–31
26. Resnik, P.: Wordnet and distributional analysis: A class-based approach to lexical discovery. In: AAAI workshop on statistically-based natural language processing techniques. (1992) 56–64
27. Tversky, A.: Features of similarity. psychological review. Tversky, Amos **84**(4) (1977) 327 American Psychological Association.
28. Sheldon, R.: A first course in probability. Pearson Education India (2002)
29. Richardson, R., Smeaton, A.: Using WordNet in a knowladge-based approach to information retrieval. Working Paper, , CA-0395, School of Computer Applications, Dublin Sity University, Ireland (1995)
30. Hughes, T., Ramage, D.: Lexical semantic relatedness with random graph walks. In: EMNLP-CoNLL. (2007) 581–589
31. Fouss, F., Pirotte, A., Renders, J.M., & Saerens, M.: Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on Knowledge and Data Engineering **19**(3) (2007) 355–369
32. Ho Lee, J., Ho Kim, M., Joon Lee, Y.: Information retrieval based on conceptual distance in is-a hierarchies. Journal of documentation **49**(2) (1993) 188–207 MCB UP Ltd.
33. Ginsberg, A.: A unified approach to automatic indexing and information retrieval. IEEE Expert: Intelligent Systems and Their Applications **8**(5) (1993) 46–56 IEEE Educational Activities Department.
34. Whan Kim, Y., Kim, J.H.: A model of knowledge based information retrieval with hierarchical concept graph. Journal of Documentation **46**(2) (1990) 113–136 MCB UP Ltd.
35. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of english: The penn treebank. Computational Linguistics **19**(2) (1993) 313–330

GOUTAM MAJUMDER
NATIONAL INSTUTUTE OF TECHNOLOGY MIZORAM,
INDIA
E-MAIL: <GOUTAM.NITA@GMAIL.COM>

PARTHA PAKRAY
NATIONAL INSTUTUTE OF TECHNOLOGY MIZORAM,
INDIA
E-MAIL: <SEE WWW.PARTHAPAKRAY.COM>

ALEXANDER GELBUKH
CIC,
INSTITUTO POLITÉCNICO NACIONAL,
MEXICO
E-MAIL: <SEE WWW.GELBUKH.COM>