Graph Based Computational Model for Computing Semantic Similarity

Jagendra Singh1,∗, Mayank Saini and Sifatullah Siddiqi

School of Computer & Systems Sciences, JNU, New Delhi 110 067, India.
e-mail: 1jagendrasngh@gmail.com; jaggu.jnu@gmail.com

Abstract. Finding semantic similarity between two natural language entities considered a challenging task in the field of natural language processing. Accuracy of presently existing semantic similarity computational methods is still very far from what humans would perceive. In this paper, we present a new approach of measuring semantic similarity/distance between concepts/words by considering all senses instead of using one most common sense of concepts in WordNet hierarchy. Our proposed approach considers not only the semantic distance between two concepts/words but also considers feature information of WordNet graph. When tested on benchmark data set of words pair similarity ratings, the proposed approach performs better than other semantic similarity computational models for ambiguous words/concepts (which has more than one sense). Proposed approach gives the highest correlation coefficient value with human similarity judgments based benchmark data set.

Keywords: WordNet, Semantic similarity, Graph theory, Transition probability, Correlation coefficient.

1. Introduction

In the field of text processing, it is often required to find similarity between two text documents. As the smallest logical unit of text is word. So ultimately there should be some means of finding similarities between words. Semantics is the most natural way of defining similarity between words. Two words are semantic similar if they have same meaning or they are relating on each other. Some background knowledge required for computing semantic similarity; ontology being most popular used for such purpose [1]. There are many potential types of relation that can consider: hierarchical (e.g. IS-A or hyponym-hypernym, part-whole, etc.), equivalence (synonyms), etc. WordNet [2] is most often used linguistic ontology for computing semantic similarity between two words.

Computing semantic similarity is a challenge because the concept of semantic similarity is not well defined as it is bases on human judgments. Therefore different methods of computing semantic similarity have been developed, each having its pros and cons. Another challenge is developing an automatic approach for finding similarity between the text documents based on semantic similarity between words present in text. Apart from many challenges in automating, the basic challenge is dealing with word sense disambiguation [13]. The main problem is polysemy, a word may have multiple senses (polysemy) and takes a particular meaning based on the sense it is commonly used. Again automatic word sense disambiguation is a challenge and there is no algorithm that can do this task with 100% accuracy. All the method developed for finding semantic similarity expects that the correct sense of word is known.

In absence of correct sense knowledge, either most often accessed sense used which may or may not be correct or a word sense disambiguation algorithm is used which again may not return the correct sense. Considering these constrains, we propose an approach that computes semantic similarity between two concepts using all senses of the concept. The proposed approach is based on transition probabilities of nodes in WordNet graph [3]. The experiment shows that the result of our proposed similarity approach is better than the other existing similarity approaches.

The rest of this paper organized as follows: In section 2, current approach of semantic similarity discussed and compared. The proposed approach described in section 3. In section 4, experiment will take place to evaluate various semantic similarity measures and results compared with human similarity judgment. Finally, conclusion and future work discussed in section 5.

∗Corresponding author
2. Related Work

WordNet [2] Taxonomy is often represented as a hierarchical structure, which can be seen as a special case of network structure. For evaluating semantic between nodes in the network, one can make use of the structural information embedded in the network. There are several ways to find the conceptual similarity of two words in a hierarchical semantic network. According to the approaches used to find semantic similarity, it can be classified into two main categories, the edge-based approach and the information content based approach.

2.1 The edge based approach

Leacock and Chodorow (Lch) [4] proposed a semantic similarity measure that incorporates edge-based approach. This similarity measure determines similarity by using the length of shortest path that connects two concepts in the WordNet taxonomy. We can express this measure as follows:

\[ \text{sim}_{Lch}(c_1, c_2) = -\log \left( \frac{\text{len}(c_1, c_2)}{2D} \right) \]  

where \( \text{len}(c_1, c_2) \) denotes the length of the shortest path between the concepts \( c_1 \) and \( c_2 \), \( D \) denotes the maximum depth of WordNet hierarchy. Basically maximum depth, which is constant, acts as a normalization factor for similarity value.

Wu and Palmer (Wup) method [10] calculates similarity between two concepts by considering the depths of both the concepts in the WordNet hierarchy, along with the depth of the least common subsumer (Lcs). Least common subsumer is the subsumer which subsumes both the concepts at lowest level in hierarchy.

\[ \text{sim}_{Wup}(c_1, c_2) = \frac{2 \times \text{depth}(\text{lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \]  

where \( \text{lcs}(c_1, c_2) \) is the lowest common subsumer of \( c_1 \) and \( c_2 \). The depth \( (c_1) \) is the length of the path from \( c_1 \) to root.

2.2 The information content based approach

According to the idea of information content based measures, the two concepts that share more information content means more similar the concepts are. If \( c \) is a common ancestor of two concepts and \( p(c) \) is frequency of \( c \), then information content (IC) of \( c \) calculated with the help of corpus [6] and given by:

\[ \text{IC}(c) = -\log p(c) \]

In a concept tree of ontology, we see that IC is decreasing as we move from leaves to the higher level in hierarchy that indicates the more generality of the concept and denotes more specificity. Resnik P. [8] proposed the following formula for calculating semantic similarity based on IC.

\[ \text{sim}(c_1, c_2) = \max_{c \in S(c_1, c_2)} (-\log p(c)) \]  

where, \( S(c_1, c_2) \) is the set of all common ancestor nodes of \( c_1 \) and \( c_2 \). That measure how similar two concepts are by using the most IC from all common ancestor.

Jiang and conrath [3] proposed a semantic similarity measure that combines edge-base and information content based approach to compute the total semantic distance between two concepts in taxonomy. This measure uses the sum of the individual distance between the nodes in the shortest path. The semantic distance between any two concepts \( c_1 \) and \( c_2 \) in the taxonomy is:

\[ \text{dist}_{jc}(c_1, c_2) = \text{IC}(c_1) + \text{IC}(c_2) - 2 \times \text{IC}(\text{lcs}(c_1, c_2)) \]  

3. Proposed Model

In WordNet, most of the concepts have more than one sense. The similarity measures discussed in section 2 take only one sense to represent the concept and ignore other senses information. Here, we propose a semantic similarity approach which considers all senses of the concepts and find more accurate semantic similarity between concepts.

The concept ward has four senses and the concept hospital has one sense in WordNet. The integrated hierarchical structure of concepts ward and hospital in WordNet taxonomy represents as a directed acyclic graph. Every oriented segment in this graph represent IS-A relation. There are four paths between concepts hospital \( (n_1 \) and ward \( n_1 \)).
Graph Based Computational Model for Computing Semantic Similarity

3.1 Concepts used in the proposed approach

Concept based graph: \( G = (N, E, R) \) is a directed acyclic graph. Where \( N \) represents the node set, \( E \) represents the set of directed edges between nodes and \( R \) represents the root of the graph.

Path: \( G = (N, E, R) \) is a concept based graph with nodes \( n_0, n_1, n_2, \ldots n_m \in N, m \geq 0 \). Directed edges \((e_1, e_2, e_3, \ldots, e_m)\) follows the condition that any \( e_j \in E(1 \leq j \leq m) \) is directed edge from \( n_{j-1} \) to \( n_j \), then the node sequence \( P = (n_0, n_1, n_2, \ldots, n_m) \) is the path from \( n_0 \) to \( n_m \).

Concept path: In graph \( G = (N, E, R) \), the directed path from \( R \) to any node \( n \in N \) is called a concept path for \( n \).

Transition probability of concept path: \( P = (n_0, n_1, n_2, \ldots n_m) \in \varphi(n_m)(n_0 = R) \) is a concept path for \( n_m \cdot p(e_j) \) represents the probability from \( n_{j-1} \) to \( n_j \). \( \prod_{j=1}^{m} p(e_j) \) is called the transition probability of \( P \).

Transition probability of the node: \( \varphi(n) \) is the concept path set for node \( n \), the transition probability for node \( n \) is the sum of transition probability of all concept paths in the concept set, that is:

\[
p(n) = \sum_{\varphi(c)} \prod_{j=1}^{m} p(e_j)
\]

A proposed semantic similarity model based on multipath transition probability discussed in next part.

3.2 Proposed semantic similarity model

Proposed semantic similarity approach based on WordNet is as follows:

\[
 sim_{proposed}(c_1, c_2) = \alpha \frac{\gamma}{dist_{jc}(c_1, c_2)} + (1 - \alpha) \left( \frac{2 \times \log p(lcs(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \right)
\]

Here, \( dist_{jc}(c_1, c_2) \) is Jiang and Conrath [3] semantic similarity measuring approach, \( \alpha \in [0, 1] \) is power weight. \( lcs(c_1, c_2) \) is the lowest common subsumer of concepts \( c_1 \) and \( c_2 \). \( p(c) = \sum_{e_j \cdot \gamma}^{p(e_j)} \) is the \( c \)’s transition probability, \( e_j \) is the directed edge from \( c_{j-1} \) to \( c_j \). \( \frac{1}{\log p(c_1) + \log p(c_2)} \) is the common information relation between two concepts.

3.3 Proposed algorithm

Input: Two concept/word pairs \( c_1 \) and \( c_2 \).

Output: Numeric value of Semantic similarity between \( c_1 \) and \( c_2 \).

Algorithm

Step 1: Input concepts \( c_1 \) and \( c_2 \).
Step 2: Concepts validation, If both concepts are present in WordNet, goto Step 3, else goto Step 8.
Step 3: Find all hyponym tree of concepts \( c_1 \) and \( c_2 \) using WordNet Taxonomy (may be more than one tree for each concept).
Step 4: Find all hyponym trees have same root or not, If root same, goto step 5, else goto step 8.
Step 5: Find nearest common hyponym ancestor node of both concepts \( c_1 \) and \( c_2 \)’s hyponym trees, which is called least common subsumer (Lcs).
Step 6: Transition probability based Similarity module, Find all paths from root to concepts \( c_1 \) and \( c_2 \) in their hyponym trees, Find information content of concept nodes \( c_1, c_2 \) and Lcs node (equation 2). Find transition probability of all paths of both concepts \( c_1 \) and \( c_2 \) (equation 1), Find semantic similarity between both concepts using proposed semantic similarity model (equation 6).
Step 7: Output numeric value of semantic similarity between concepts \( c_1 \) and \( c_2 \).

Step 8: Stop.

Here, we use concepts hospital \((n_{17})\) and ward \((n_{18})\) to measure semantic similarity as an example shown in figure 1. In figure 1, Concept hospital have one sense, so it has one path from root \((n_0)\) to hospital \((n_{17})\) and concept ward has four senses, so it has four paths from root \((n_0)\) to ward \((n_{18})\) in WordNet lexical hierarchy. In order to calculate semantic similarity between hospital and ward, we use all senses of both concepts. Here, \( c_1 = \text{Hospital} \), \( c_2 = \text{Ward} \).

**Based on WordNet**

\[
\text{Lcs}(\text{Hospital, Ward}) = \text{Object}; \quad \text{IC}(\text{Hospital}) = 0.6957; \quad \text{IC}(\text{Ward}) = 0.8772; \quad \text{IC}(\text{Object}) = 0.0611; \quad \text{dist}_{jc}(\text{Hospital, Ward}) = 1.4507 \text{ using (equation 6, equation 3)}. \quad \text{Now } \quad p(\text{Hospital}) = \sum_{x=0}^{\#} p(e_j) = p(n_0) \cdot p(n_1) \cdot p(n_2) \cdot p(n_3) \cdot p(n_4) \cdot p(n_{12}) \cdot p(n_{16}). \quad \text{Such that, } \quad \log p(\text{Hospital}) = 52.97, \quad \text{and } \quad p(\text{Ward}) = \sum_{y=0}^{\#} p(e_j) = p(n_0) \cdot p(n_1) \cdot p(n_2) \cdot p(n_3) \cdot p(n_4) \cdot p(n_{12}) \cdot p(n_{15}) + p(n_0) \cdot p(n_1) \cdot p(n_3) \cdot p(n_6) \cdot p(n_{12}) \cdot p(n_{13}) \cdot p(n_{16}) + p(n_0) \cdot p(n_1) \cdot p(n_3) \cdot p(n_4) \cdot p(n_6) \cdot p(n_{12}) \cdot p(n_{13}) \cdot p(n_{15}). \quad \text{Such that, } \quad \log p(\text{Ward}) = 164. \quad \text{Now, } \quad P(\text{Lcs}(\text{Hospital, Ward})) = p(\text{Object}) = p(n_0) \cdot p(n_1) + p(n_0) \cdot p(n_2) \cdot p(n_3) \cdot p(n_{12}) \cdot p(n_{13}) \cdot p(n_{16}). \quad \text{Such that, } \quad \log p(\text{Object}) = 72.20 \quad \text{By proposed semantic similarity model (equation 6) } \text{Sim}(\text{Hospital, Ward}) = 1.2249.

Here, \( IC(c) = 1 = \frac{-\log(\text{hypo}(c) + 1)}{\log(\text{maxWN})} \) is used to calculate information content of concept \( c \) [9,11]. hypo \((c)\) is the total number of hyponym of concept \( c \) in WordNet taxonomy and \( \text{maxWN} \) is the maximum number of nouns in WordNet 2.0 (which is equal to 79689) [11]. We set \( \alpha = 0.6 \) as power weight and \( \gamma = 20 \) to make Jiang and Conrath’s similarity between \((0.5–1.0)\) from \((0–19)\) in our experimental results.

4. Experiment and Results

In this section, we present our experiment to find semantic similarity between concepts/words. In order to evaluate the efficiency of our proposed method, we compare our results with the four methods discussed in section 2. When comparing two sets of ratings, we are interested in the strength of the linear association between two quantitative variables, so we follow the comparison of results by the correlation coefficient of each computational measure with the human ratings, see table 2.

For the experiment we used Miller and Charles [7] benchmark dataset of 30 noun word pairs in which human judgment involved. In order to make fair comparison we used a software package developed by Ted Pederson [12], which implement semantic similarity measures described by Leacock and Chodorow, Jiang and Conrath, Resnik, Lin, Wu and Palmer, etc to calculate semantic similarity between words/concepts.

The results of semantic similarity measures for word pairs are shown in table 1 (given below). We choose suitable values of \( \alpha \) and \( \beta \). Table 1, show the semantic similarity value between 30 word pairs using different semantic similarity approaches with proposed similarity approach.

The relationship between power weight and correlation coefficient [5] used in our proposed approach shown in following graph.
Table 1. Semantic similarity values of different similarity models with human rating for word pairs.

<table>
<thead>
<tr>
<th>Words Pairs</th>
<th>Humans Ratings(0-4)</th>
<th>$\text{Sim}_{L\text{ch}}$ (0 - 4)</th>
<th>$\text{Sim}_{W\text{up}}$ (0 - 1)</th>
<th>$\text{Sim}_{R}$ (0 - 16)</th>
<th>$\text{Sim}_{J}$ (0 - 1)</th>
<th>$\text{Sim}_{\text{proposed}}$ (0 - 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>automobile</td>
<td>car</td>
<td>3.92</td>
<td>3.68</td>
<td>1.0</td>
<td>6.31</td>
<td>0.26</td>
</tr>
<tr>
<td>gem</td>
<td>Jewel</td>
<td>3.84</td>
<td>1.60</td>
<td>0.63</td>
<td>2.49</td>
<td>0.06</td>
</tr>
<tr>
<td>journey</td>
<td>voyage</td>
<td>3.84</td>
<td>2.30</td>
<td>0.86</td>
<td>6.79</td>
<td>0.35</td>
</tr>
<tr>
<td>boy</td>
<td>lad</td>
<td>3.76</td>
<td>2.59</td>
<td>0.85</td>
<td>4.64</td>
<td>0.19</td>
</tr>
<tr>
<td>coast</td>
<td>shore</td>
<td>3.70</td>
<td>2.99</td>
<td>0.92</td>
<td>8.10</td>
<td>1.0</td>
</tr>
<tr>
<td>asylum</td>
<td>madhouse</td>
<td>3.61</td>
<td>1.60</td>
<td>0.66</td>
<td>3.94</td>
<td>0.0</td>
</tr>
<tr>
<td>magician</td>
<td>wizard</td>
<td>3.50</td>
<td>1.89</td>
<td>0.66</td>
<td>1.90</td>
<td>0.63</td>
</tr>
<tr>
<td>midday</td>
<td>noon</td>
<td>3.42</td>
<td>3.68</td>
<td>1.0</td>
<td>9.56</td>
<td>1.0</td>
</tr>
<tr>
<td>furnace</td>
<td>stove</td>
<td>3.11</td>
<td>1.12</td>
<td>0.5</td>
<td>2.49</td>
<td>0.05</td>
</tr>
<tr>
<td>food</td>
<td>fruit</td>
<td>3.08</td>
<td>1.29</td>
<td>0.37</td>
<td>0.61</td>
<td>0.09</td>
</tr>
<tr>
<td>bird</td>
<td>cock</td>
<td>3.05</td>
<td>0.91</td>
<td>0.28</td>
<td>0.61</td>
<td>0.0</td>
</tr>
<tr>
<td>bird</td>
<td>crane</td>
<td>2.97</td>
<td>1.49</td>
<td>0.63</td>
<td>1.82</td>
<td>0.0</td>
</tr>
<tr>
<td>implement</td>
<td>tool</td>
<td>2.95</td>
<td>2.99</td>
<td>0.94</td>
<td>6.31</td>
<td>0.85</td>
</tr>
<tr>
<td>brother</td>
<td>monk</td>
<td>2.82</td>
<td>1.60</td>
<td>0.58</td>
<td>1.90</td>
<td>0.68</td>
</tr>
<tr>
<td>crane</td>
<td>implement</td>
<td>1.68</td>
<td>1.38</td>
<td>0.52</td>
<td>1.36</td>
<td>0.0</td>
</tr>
<tr>
<td>brother</td>
<td>lad</td>
<td>1.66</td>
<td>1.74</td>
<td>0.62</td>
<td>1.90</td>
<td>0.08</td>
</tr>
<tr>
<td>car</td>
<td>journey</td>
<td>1.16</td>
<td>0.69</td>
<td>0.17</td>
<td>0.07</td>
<td>0.29</td>
</tr>
<tr>
<td>monk</td>
<td>oracle</td>
<td>1.10</td>
<td>1.60</td>
<td>0.58</td>
<td>1.91</td>
<td>0.06</td>
</tr>
<tr>
<td>cemetery</td>
<td>woodland</td>
<td>0.95</td>
<td>1.49</td>
<td>0.50</td>
<td>1.16</td>
<td>0.05</td>
</tr>
<tr>
<td>food</td>
<td>rooster</td>
<td>0.89</td>
<td>0.91</td>
<td>0.28</td>
<td>0.61</td>
<td>0.07</td>
</tr>
<tr>
<td>coast</td>
<td>hill</td>
<td>0.87</td>
<td>2.07</td>
<td>0.71</td>
<td>6.31</td>
<td>0.22</td>
</tr>
<tr>
<td>forest</td>
<td>graveyard</td>
<td>0.84</td>
<td>1.04</td>
<td>0.23</td>
<td>0.0</td>
<td>0.06</td>
</tr>
<tr>
<td>shore</td>
<td>woodland</td>
<td>0.63</td>
<td>2.07</td>
<td>0.66</td>
<td>1.16</td>
<td>0.06</td>
</tr>
<tr>
<td>monk</td>
<td>slave</td>
<td>0.55</td>
<td>2.07</td>
<td>0.71</td>
<td>1.90</td>
<td>0.06</td>
</tr>
<tr>
<td>coast</td>
<td>forest</td>
<td>0.42</td>
<td>1.29</td>
<td>0.28</td>
<td>0.0</td>
<td>0.06</td>
</tr>
<tr>
<td>lad</td>
<td>wizard</td>
<td>0.42</td>
<td>2.07</td>
<td>0.71</td>
<td>1.09</td>
<td>0.07</td>
</tr>
<tr>
<td>chord</td>
<td>smile</td>
<td>0.13</td>
<td>1.29</td>
<td>0.37</td>
<td>0.78</td>
<td>0.06</td>
</tr>
<tr>
<td>glass</td>
<td>magician</td>
<td>0.11</td>
<td>1.49</td>
<td>0.42</td>
<td>0.61</td>
<td>0.05</td>
</tr>
<tr>
<td>rooster</td>
<td>voyage</td>
<td>0.08</td>
<td>0.51</td>
<td>0.14</td>
<td>0.0</td>
<td>0.05</td>
</tr>
<tr>
<td>noon</td>
<td>string</td>
<td>0.08</td>
<td>0.85</td>
<td>0.20</td>
<td>0.0</td>
<td>0.05</td>
</tr>
</tbody>
</table>

This is clear from the above graph that correlation coefficient ($\sigma$) of proposed approach reach maximum when power weight value reach ($\propto= 0.6$). Lch and Wup approach belongs to edge-based approach. Resnik, Jiang and Conrath belongs to information content based approach. All these method consider only one sense of the word. But in our proposed approach, we consider all senses of the word which described the reality more accurately. Below table shows the Correlation coefficient between different semantic similarity approaches and human judgment.

It observed from the correlation table 2 that the overall result of proposed approach is better than other similarity approach. The relationship between human judgment similarity rating and semantic similarity measures similarity value discussed in table 1 shown in below correlation graph.
In this graph’s figure, we looks that in our proposed model, the distribution of semantic similarity of concept pairs is continuous, more centralized, especially there is no fault zone as we see in other models. And table 2 of correlation coefficient indicates better performance of our proposed model.

On analyzing results we can make following observation

- In our experiment, the proposed approach based on information content and transition probability are performing better than other edge based and information content based measure.
- Lch similarity measures, which is edge based measure based on fixed depth is giving better results than wup, which is edge based measure on variable depth i.e. it takes into account depth of both the words/concepts.
This means, there is a scope of improvement in modifying Wup method, where in addition to using depth based information, some other graph based properties can be used.

- The information content based measure has possibility to improve by using a better corpus or by incorporates some other graph properties.

5. Conclusion and Future Work

Overall purpose of this work was to develop a lexical ontology based module for finding semantic similarity between two concept pairs considering all senses of concepts. This module can incorporate and utilize the potentially valuable information contained in WordNet ontology. In this paper we have proposed algorithm and developed java modules (for English text) for finding semantic similarity between two concepts. Four popularly semantic similarity methods (Resnik, Lch, Wup and Jiang & conrath) were implemented. We experimented with these Semantic Similarity Measures for English concepts and evaluated their performance on Benchmark data set. Based on comparison between our experimental results and standard results with the help of correlation coefficient, we find that our proposed semantic similarity measure is performing well among four semantic similarity measures in our experiment.

As discussed earlier, there is a lot of scope for improving and developing new semantic similarity methods. We will attempt to evaluate our proposed model on a large dataset and for specific application, such as text summarization, information extraction, information retrieval etc. In future we may try to improve these methods. One of the major limitations we have observed is that all measures used by us are using only IS-A relation and information content(IC) of a concept to calculate similarity. But many factors/properties of lexical ontology are not taken into account yet, i.e. properties of a concept and instance of a concept. Making use of these factors we may able to design better semantic similarity measure.

References