Measuring of Similarity between Pair of Words Using Word Net

Atul Gupta [1], Kalpana Sharma[1] and Krishan Kumar Goel[2]

[1]*Department of Computer Science and Engineering, Bhagwant University, Ajmer, India*
[2]*Department of Computer Application, Raja Balwant Singh Management Technical Campus, Agra, India*

Abstract

In the current era, the digital data size increased enormously and available abundantly. Retrieving the relevant and accurate information from the available data still a big challenge. In this paper we are finding the similarity of noun-noun pairs and verb-verb pairs using Word-Net as corpus. Computation of similarity between noun-noun pairs in a sentence using different semantic algorithm computed and analysed.it has been observed that computation of similarity between verb-pair is found to be not as easy as computation of similarity between noun-pair. There are two challenges observed during the experimentation of this work. The first one is no standard data set available for verb pair and second is no exact hierarchy of verb of available in word-net.

Keywords

WordNet, Similarity Measure, Semantic Similarity, IS-A relationship

1. Introduction

Internet Searching has been integral part of life. There are lot of search engines available, but it has still some unsolved challenges like not able to provide accurate and exact search result. For example, if someone searches "Lincoln a Car " then it will be providing Car brand as well as it will provide output about Abraham Lincoln also. This search is not about the president of America President Abraham Lincoln, but the search engine will display these results also.

In the given example it is the Car and the Name Lincoln have been identified as noun-noun pair. So semantically it should relate the two nouns together and result will be retrieved as per the context given. Computation of similarity between word pair (noun-noun/ verb-verb) and sentence pair is still a huge problem for the researcher who work in the field of search engine, gene prioritization and NLP. Measuring of similarity between words is possible only in fixed domain for example: medical domain, engineering domain etc. Computing the similarity between noun pairs and verb pairs is done by lexical database. Words are connected in the form of lexical chain in lexical database i.e., Lucknow→City→Capital→Uttar Pradesh. Different semantic measure algorithms have been developed to compute the semantic closeness in the pair of words using the lexically connected database i.e., Word-Net. Words are present in hierarchal form in Word-Net. Various approaches have been implemented previously which uses lexical database as Word-Net. George A. Miller [1] a psychology professor of Princeton university developed Word-Net in 1985.

Words are arranged in Word-Net corpus in the synonymous relationship called as Synsets. In the Word-Net, words are organized in the form of Synset. There are 207016 word-pairs presents in compressed form. Word-Net contain different type of semantic relationship like synonym, antonym, hyponyms and meronyms in noun, verb, adverb and adjective.

ACI'22: Workshop on Advances in Computation Intelligence, its Concepts & Applications at ISIC 2022, May 17-19, Savannah, United States
EMAIL: atulpsit3883@gmail.com (A. 1); kalpanasharma56@gmail.com (A. 2); kkgoyal@gmail.com (A. 3)

Table 1
Noun relationship in Word-Net (Corpus)

Relationship	Definition	Example
Hypernym	concept to subordinate relationship	word1="breakfast"→word2="meal"
Hyponym	concept to sub-type relationship	word1="plant"→word2="tree"
Part-of	part to whole relationship	word1="course"→word2="meal"
Has-part	whole to part relationship	word1="table"→word2="leg"
Antonym	opposite relationship	word1="leader"→word2="follower"
Hypernym	concept to subordinate relationship	word1="breakfast"→word2="meal"

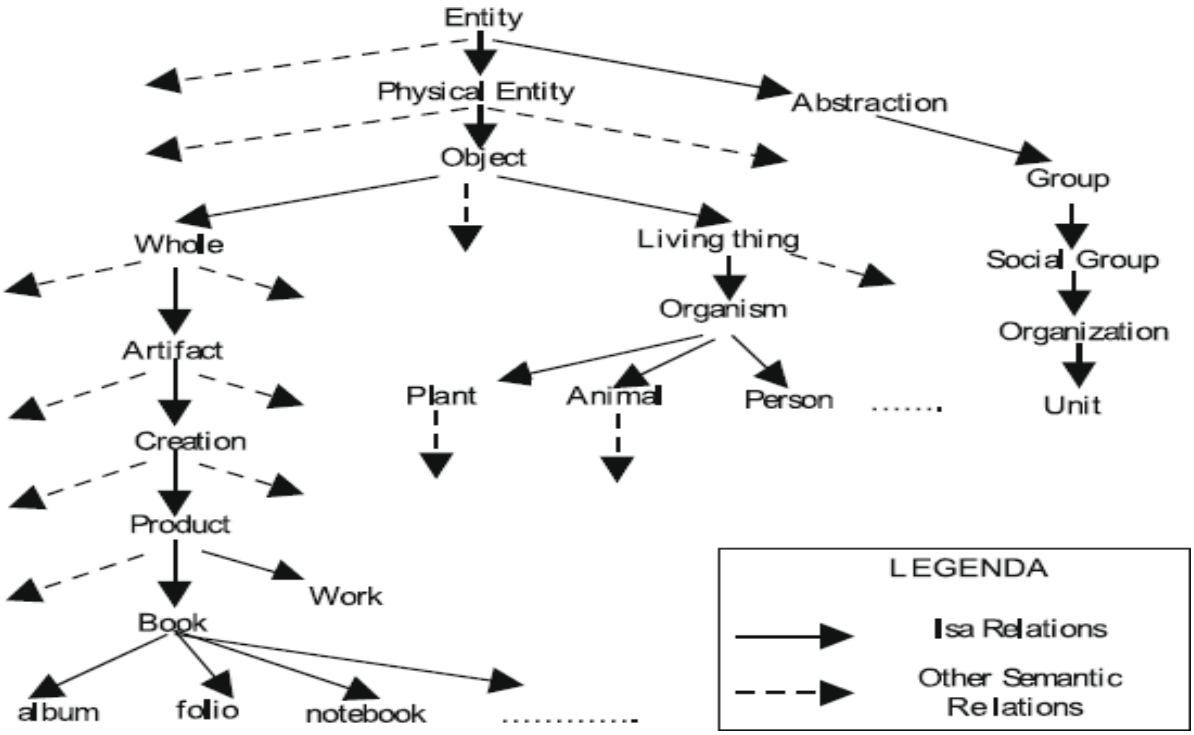

Figure 1: A Fragment of WordNet taken from [3] for illustration

Semantic similarity is classified in different cateogories-:

1. Method based on word co-occurrence
2. Method based on lexically connected database
3. Web-Search-Engine Method.

The Co-occurrence approach[2] formalizes the concept concerning information retrieval. In the word co-occurrence methodolgy, there is a word-list and for each word in the word list a meaningful word is connected.The query is retrieved by creating a vector. Ordering and context of this particular search query is fully unmeasured. So it is a major drawback of this methodolgy. In the lexical database methodology[3], similarity computation is done by the pre-defined Word-Net hierarchy, which is arranged in tree-like structure[3]. Web search engine methodolgy compute the similarity but sometimes

the word have opposite meaning which may occurred in the same webpage. This influenced the calculated similarity value adversly. This methodolgy is developed by Google-Similarity-Distane[4].

2. Overview Of Different Similarity Measure Technique

The computation of similarity between two word pair i.e noun-noun pair/ verb-verb pair is done by using Word-Net ontology. Word-Net is developed by professor G.A Miller[1] and it is managed by the laboratory of Conginitive science in princeton university. The ontology contain three databases noun, verb and adverb-adjective. In this ontology, words are organized in the form of Synsets.This paper is focussed on IS-A relationship between noun pair and verb pair. There are several hierarchy present in the Word-Net and all the hierarchy are subsumes in a common root node. Similarity approach are classified in different forms:

(a) Similarity calculation by distance based methodology

(b) Similarity calculation by Information Content based methodology.

(c) Similarity calculation by Feature based methodogy.

2.1. Similarity calculation by distance-based methodology

In this approach similarity is computed by measuring the distance between two words. This is also called Edge-Counting based methodology. Pair of words concerning Path Length is calculated for measuring the similarity among group of words. Similarity score measured by this approach is in discrete form, so there is applying normalization. Various path-based algorithms have been developed by Leacock-Chodorow[7] and Wu and palmer[9].

2.1.1. Leacock-Chodorow Similarity Approach

The Approach of LCH [7] is based on shortest path length. In this approach computation of similarity is done by finding the path length. It is the shortest path between noun-pair in the Word-Net IS-A relationship. The shortest path is defined as there are less no of intermediate node between two words. The shortest value retrieved by this is scaled up by the depth factor D where calculation of depth is done by the longest path from the root to leaf in the hierarchy of Word-Net. The calculation of similarity is done by.

$$Similarity_{LCH} = -log\left\{\frac{(minimum(length(w_1, w_2)))}{2 * D}\right\} \quad (1)$$

where w1 denotes first word and w2 denotes second word. minimum(length(w1,w2)) denotes the minimum path length between word pair w1 and w2. Depth factor D is the maximum depth from root to leaf in the Word-Net.

LCH approach is easy. Computation of similarity between word pair is done by counting the number of links between word-pairs.

2.1.2. Wu and Palmer Similarity Approach

The Wu and Palmer Similarity Approach[9] focuses on path length among word-pair. This is based on most specific common predecessor node.It is called as Lowest Common Subsumer node(LCS).The Similarity between two words in a IS-A relationship in Word-net ontology is computed in this manner.

This method perform well in verb ontology and different part of speeches where words are arranged in hierarichal structure. The calculation of similarity is done by:

$$\text{Similarity}_{\text{wu\&palmer}}(\text{w1}, \text{w2}) = 2 * \frac{C}{A} + B + 2 * C \quad (2)$$

Where A and B define as count of IS-A node between word w1 & w2 to the most common ancestor node. C is defined as depth. It is computed from the word present in the Word-Net to the root in the Word-Net. The Approach of Wu and Palmer is based on the lowest common subsumer of two words. The value of similarity between words never becomes 0. In the Fig2.1. Comparison of similarity between words is determine by LCH and Wu &Palmer based.

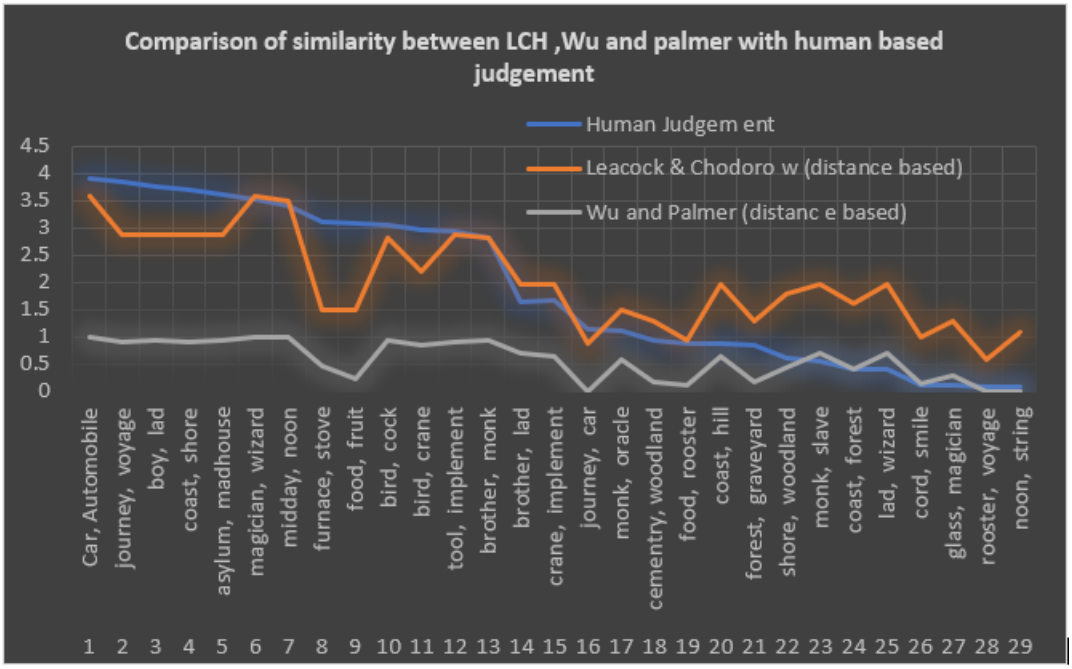

Figure 2: Comparison of similarity between LCH ,Wu and palmer with human based judgement

2.2. Similarity calculation by Information Content

It deals with the similarity characteristic is calculated on the basis of information content. The content of Information concerning the word is computed by the frequency of the word in the Word-net ontology. The characteristic of frequency concerning the word is computed by the probability of occurrence of the word. Information content of the word is computed as:

$$\text{Information Content(IC)} = -log(p(c)) \quad (3)$$

2.2.1. Resnik Similarity Approach

Resnik [10] similarity approach is based on the value of information content(IC) of the word.In this approach Similarity value calculation is done by how much information is shared between words w1 and w2. If the information shared between words are more than similarity value is high, otherwise low. The calculation of similarity is done by:

$$Similarity_{RES}(w1,w2) = IC(lcs(w1,w2) \quad (4)$$

This approach of computing similarity work on verb along with noun. The justification for this that both part of speech in structured in hierarichal manner[5,6].The value computed by resnik simlarity is always greater than 0.The Information Content value give better result than path-based approaches.

2.2.2. Similarity Approach of Jiang & Conrath

Jiang and Conrath [11] approach depends on the information content of the LCS between words and semantic distance between word pairs. The calculation of similarity is done by:

$$\boldsymbol{Distance_{Jiang\&Conrath}(W_1, W_2) = ICvalue(W_1) + ICvalue(W_2) - 2 * ICvalue(lcs(W_1, W_2))} \quad (5)$$

$$\boldsymbol{Similarity_{Jiang\ and\ Conrath} = \frac{1}{Distance_{Jiang\&Conrath}(W_1, W_2)}} \quad (6)$$

Where Distancejiang and conrath find the value of dissimilarity between words. The measure of values (low) indicate more similar words and high value indicate least similar words.

Jiang and Conrath approach is same as Resnik's approach. Similarity find by this approach is based on commonality between pair of words w1 and w2 and IC value of the words. There is special case need to handle when the value is 0.

2.2.3. Lin Similarity Approach

Lin similarity approach[12] is based on the information that is shared by the word pairs to the summation of Information content value of the the word pairs. The concept of Lin's is based on the commonality in word-pairs to the information content value that described them completely.

$$\boldsymbol{Sim_{Lin} = 2 * IC(lcs(W_1, W_2,)) * IC(W_1) + IC(W_2)} \quad (7)$$

Commonality means the information that is shared by the common ancestor node of the word w1 and w2 and the total amount of information described completely by the word w1 and w2. Lin computes the similarity between 0 and 1. 0 denotes the low similarity that mean two words are of different context and 1 denote the high similarity that mean two word are same or compared with itself. The similarity is calculated by:

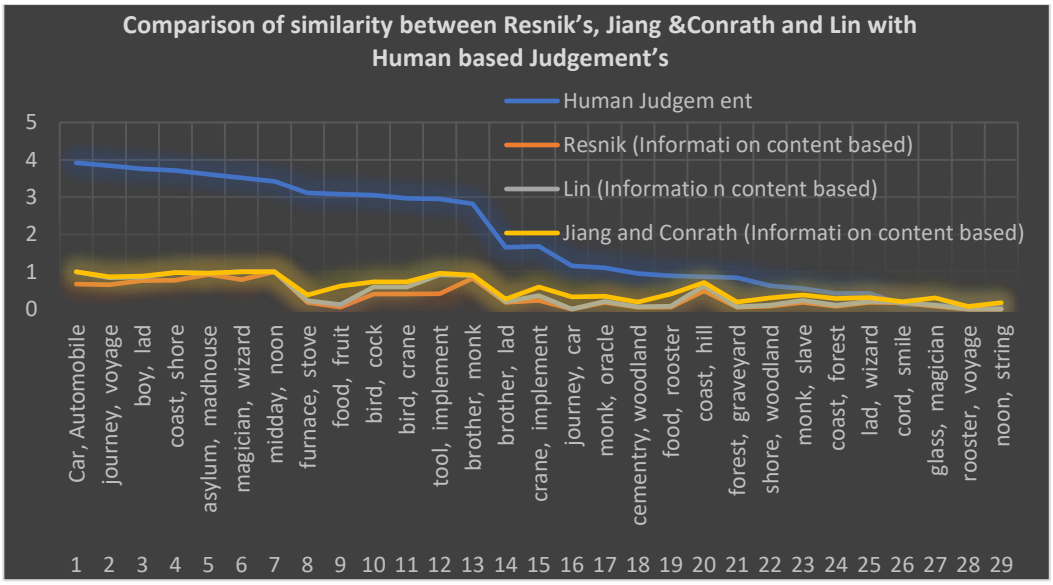

Figure 3: Comparison of similarity between Resnik's, Jiang &Conrath and Lin with Human based Judgement's

2.3. Similarity Computation by Feature based methodology

This similarity is based on features. Similarity value is high if the two words share the more common features and similarity value is low if the words have some unique features.

2.3.1. Tversky's Similarity Approach

Tversky's similarity approach[13] is based on features. If the two word pairs w1 and w2 have more common features then similarity is high, and if the word pairs have indegenious features then the meaure of similarity values will be low. The Measure of Similarity between two words is based on how much the two words shared the common feature[8], the unique feature present in word w1 but not present in word w2 , and the unique feature present in word w2 but not present in word w1. The similarity is calculated by:

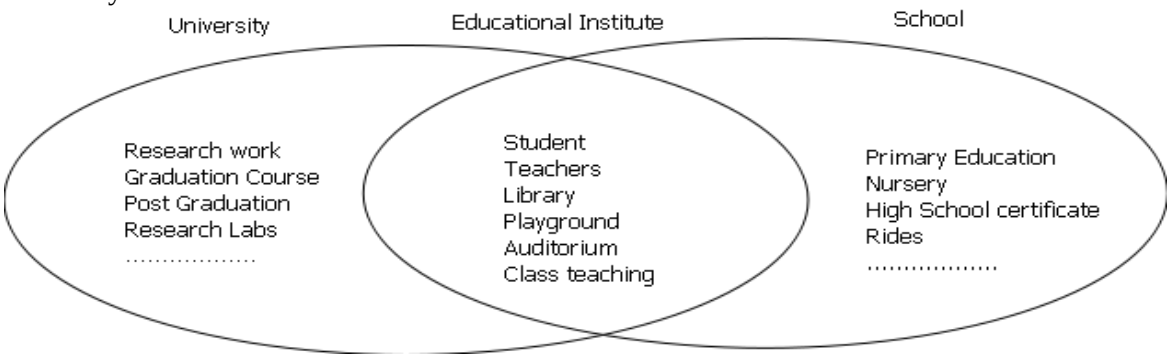

Figure 4: Tversky Feature based Approach

$$\boldsymbol{Similarity_{tversky's}} = \boldsymbol{a.G(f(w1) \cap f(w2)) - \frac{b\,G(f(w1)}{f(w2)} - \frac{c\,G(f(w2)}{f(w1)}} \qquad (8)$$

Where $\frac{f(w_1)}{f(w_2)}$ means the unique feature of w_1 but not present in w_2

$\frac{f(w_2)}{f(w_1)}$means the unique feature of w2 but not present in w1. f(w1)intersection f(w2) means feature similar in words.

2.3.2. The Similarity Approach of Piarro

Piarro approach[14] focuses on the feature based approach developed by Tversky's . This approach explored the Tversky's feature based approach into information content(IC) domain.
The function $G(x(w_1)\ intersection\ x(w_2)$ is equivalent to IC(lcs(w1,w2)).
$\frac{x(w_1)}{x(w_2)}$ is equivalent to IC(w1)-IC(lcs(w1,w2)) , $\frac{x(w_2)}{x(w_1)}$is equivalent to IC(w2)-IC(lcs(w1,w2)).

$$\mathbf{Similarity_{piaroo}} = \begin{cases} \mathbf{3} * \boldsymbol{ICvalue(lcs(w1,w2))} - \boldsymbol{ICvalue(w1)} - \boldsymbol{ICvalue(w2)} & \boldsymbol{if(w1 \neq w2)} \\ \mathbf{0} & \boldsymbol{if(w1 = w2)} \end{cases} \quad (9)$$

Where:
ICvalue(lcs(w1,w2)) denotes Information content value of the subsumer.
ICvalue(w1) denote Information-content value of the word w1.
ICvalue(w2) denote the Information -content value of the word w2.

3. Comparison of Various Similarity Technique

On Comparison of two distance-based similarity approach i.e., Leacock and Chodorow and Wu and Palmer and three Information content-based approach similarity approach i.e., Resnik's and Lin etc. The result shows that similarity score focuses on the synonym of the words. Rubenstein [15] selected 51 word-pair from human based judgement from set of 65 word-pairs. Rating of word-pair is given in the range from 0.0 to 4.0. The rating 0 means the words are semantically unrelated and rating 4.0 means words are highly related. Miller and Charles [16] selected 30 set of word-pair from Rubenstein and Goodenough [15] 65 set of word-pair. Miller's divide 30 word-pair into three sets i.e., 10 word-pair in each set. In the first set of 10 word-pair having rating between 3.0 to 4.0 are high similarity value words , next set of 10 word-pair having rating between 1.0 to 3.0 are intermediate similarity value words and last set of 10-word pair having rating between 0.0 to 1.0 are low similarity value words. In this paper for similarity calculation Miller and Charles 30 word-pair are taken for computation of similarity. In the Fig 5 various similarity-based algorithm as shown below.

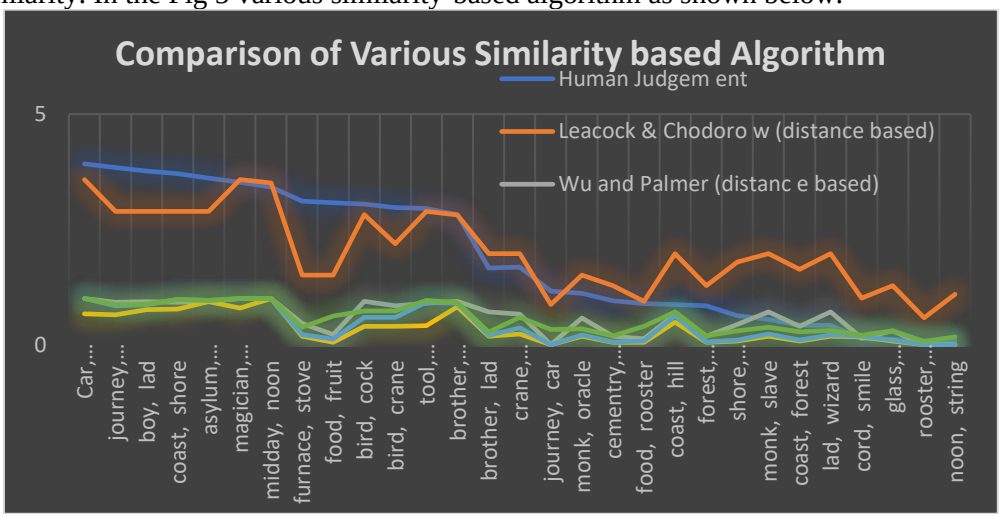

Figure 5: Comparison of Various Similarity based Algorithm

4. Results & Discussion

The result shows that on the implementation of the various edge counting methodology like LCH, Wu & Palmer and Information content methodology like Resnik's and Jiang etc. The information content methodolgies gives better correlation than all the edge counting methodolgies.Comparison of all the similarity approach is shown in Fig 6 given below:

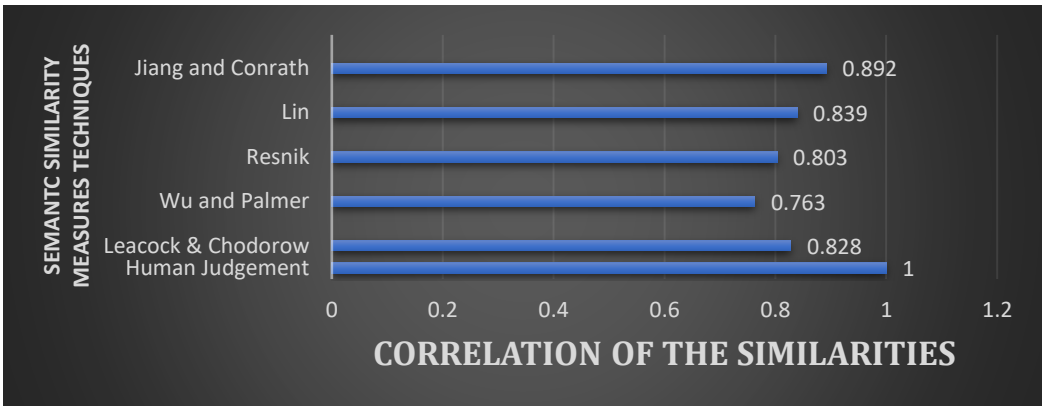

Figure 6: Semantic similarity measure vs Correlation

In the fig 6 it is observed that correlation value of jiang and conrath is better than all the other information content approaches like Resnik and Lin . The similarity value of jiang and conrath is based on commonality between words w1 and w2. and the IC-value of the words that describe them completely. The correation value of Jinag & Conrath is 0.892 on testing with Miller and Charle's 30 word-pair.

5. Conclusion

Similarity between word-pairs is one of the emerging concept in the field of artificial intelligence, machine learning and genes priortization.Calculation of similarity between

Word-pair is done by various approaches like distance based , information content based and feature based. All the approaches uses ontology of speific domain to find similarity.The Jiang and Conrath provides better result than other approaches. It is based on information and It has been seen that the similarity increases with the increase in the heirarchy of WordNet depth. So depth feature can be taken into account to find the similarity between words. The future task is to develop the approach which compute the simlarity between two pair of sentence

6. References

[1] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. "Introduction to WordNet: An on-line lexical database." *International journal of lexicography* 3, no. 4 (1990): 235-244.

[2] Boyce, Bert R., Bert R. Boyce, Charles T. Meadow, Donald H. Kraft, Donald H. Kraft, and Charles T. Meadow. *Text information retrieval systems*. Elsevier, 2017.

[3] Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38, no. 11 (1995): 39-41.

[4] Cilibrasi, Rudi L., and Paul MB Vitanyi. "The google similarity distance." *IEEE Transactions on knowledge and data engineering* 19, no. 3 (2007): 370-383.
[5] Gupta, Atul, and Krishan Kumar Goyal. "Classification of Semantic Similarity Technique between Word Pairs using Word Net."
[6] Goyal, Krishan Kumar. "Computation of Verb Similarity." *Design Engineering* (2021): 4127-4140.
[7] Leacock, Claudia, and Martin Chodorow. "Combining local context and WordNet similarity for word sense identification." *WordNet: An electronic lexical database* 49, no. 2 (1998): 265-283.
[8] Gupta, Atul, and Dharamveer Kr Yadav. "Semantic similarity measure using information content approach with depth for similarity calculation." (2014).
[9] Wu, Zhibiao, and Martha Palmer. "Verb semantics and lexical selection." *arXiv preprint cmp-lg/9406033* (1994).
[10] Resnik, Philip. "Using information content to evaluate semantic similarity in a taxonomy." *arXiv preprint cmp-lg/9511007* (1995).
[11] Jiang, Jay J., and David W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy." *arXiv preprint cmp-lg/9709008* (1997).
[12] Li, Yuhua, Zuhair A. Bandar, and David McLean. "An approach for measuring semantic similarity between words using multiple information sources." *IEEE Transactions on knowledge and data engineering* 15, no. 4 (2003): 871-882.
[13] Tversky, Amos. "Features of similarity." *Psychological review* 84, no. 4 (1977): 327.
[14] Pirró, Giuseppe. "A semantic similarity metric combining features and intrinsic information content." *Data & Knowledge Engineering* 68, no. 11 (2009): 1289-1308.
[15] Rubenstein, Herbert, and John B. Goodenough. "Contextual correlates of synonymy." *Communications of the ACM* 8, no. 10 (1965): 627-633.
[16] Miller, George A., and Walter G. Charles. "Contextual correlates of semantic similarity." *Language and cognitive processes* 6, no. 1 (1991): 1-2