Conceptual Semantic Evaluation Metric Using Taxonomy

İlknur Dönmez
Scientific and Technological Research Council of Türkiye
TÜBİTAK BİLGEM
Kocaeli, Turkey
email: ilknur.donmez@tubitak.gov.tr

Nur Bengisu Çam
Scientific and Technological Research Council of Türkiye
TÜBİTAK BİLGEM
Kocaeli, Turkey
email: nur.cam@tubitak.gov.tr

***Abstract*— The conceptual method is an important technique for calculating semantic similarity. In this study, we propose a taxonomy-based formula for calculating the conceptual similarity of sentences. The coefficients in the formula calculate how similar the noun words that make up the sentence ("verbs" and "adjectives" are also included) are to their most similar conjugates in the other sentence by considering the distance of these two words from their common ancestor and the position of the common ancestor in the ontology tree. We test our proposed metric in the English Semantic Textual Similarity (STS) benchmark dataset for semantic similarity. Although the labels of the dataset were not generated specifically for conceptual similarity, we were able to achieve 77 % accuracy in determining similar sentences using our proposed formula (which uses only noun types).**

Keywords-Similarity measures; word alignment; taxonomy; conceptual similarity, sentence similarity.

I. INTRODUCTION

In Artificial Intelligence (AI) and cognitive science, semantic similarity has become an established area of research to evaluate the strength of the semantic relationship between objects (such as words and documents). In recent years, a number of ontology-based semantic similarity metrics have been developed because they can mimic human cognitive functions. Among them, techniques based on the intrinsic information concepts have shown significant association with human evaluation [1].

According to Pirró and Euzenat [2], the scientific community divides the concepts of semantic measures into two main categories: Semantic Similarity (SS), which considers taxonomic relations such as "is-a" between two entities, and Semantic Relatedness (SR), which considers non-taxonomic relations between two entities (e.g., "cause-effect" and other associative relations such as fish lives-in water, where "lives-in" associates fish and water).

The semantic measure can be used in a variety of situations, e.g. in estimating similarity between documents [3], ontology-based text clustering [4][5], text summarization [6], entity disambiguation [7], developing recommender systems [8], semantic annotation [9], ontology merging [10], ontology segment matching [11], information retrieval [12], personalized support [13]-[15], and the graph editor similarity search problem [16], etc. Another important area is medical applications, which include automatic retrieval of patient records and medical documents [17]- [19].

The focus of this study is on semantic similarity, i.e., the "is-a" type relation between entities and ontologies is used as semantic evidence. The term "ontology" refers to any structure, such as a thesaurus, taxonomy, or other classification system, that formalizes knowledge without limiting its applicability.

Conceptual similarity comparison is an evaluation done by the human mind in order to understand semantics. The problem is to find the relationship between different concepts. In our study, after representing the sentence with its noun, verb, and adjective contents, we propose a semantic similarity metric to calculate the similarity distance between different sentences. We have made our codes publicly available on GitHub to ensure reproducibility and support future research [20].

The rest of the article is as follows: Section II contains related work on semantic similarity. Section III gives a general idea of the semantic representation of a sentence. Section IV introduces the novel similarity metric. Section V provides information about the dataset used and the evaluation results on this dataset. Section VI contains the final considerations of the metric and the results.

II. RELATED WORK

Semantic similarity of sentences has always been a popular research topic. In earlier times, methods evolved from looking at sentences word by word as a distinguishing feature to using grammatical rules to represent sentences [21]. After the creation of WordNet [22], a lexical database structured by semantic relations, ontology has been used by many researchers to compute the semantic similarity between words [23]- [28]. Jiang and Conrath measured the similarity of words by combining the taxonomy with the statistical information of the given corpus [24]. Seco et al. proposed to use WordNet for extracting the Information Content (IC) for computing the semantic similarity of words [25]. Yang and Powers proposed two different edge-based search approaches for similarity computation using WordNet [26]. Liu et al. computed the similarity between words by using the shortest part between words and the depth information from WordNet [27]. Similarly, Zhou et al. used the path length and IC value from WordNet [28]. They also compared their results with those of other authors, including Jiang and Conrath.

So far, we have mentioned the various approaches to calculating similarity between words. However, there are other studies that compute sentence similarities rather than

 ISBN: 978-1-68558-108-4

word similarities [29]- [33]. Sravanthi and Srinivasu analyzed the existing methods for computing sentence similarity and applied feature selection techniques for further investigation [29]. Selvarasa et al. used knowledge-based and corpus-based methods to measure sentence similarity in Tamil language [30]. Jeyaraj and Kasthurirathna proposed a multilayer semantic similarity network with the different number of layers and tested it on the SemEval [31] dataset [32]. Lee proposed a new approach for computing similarity between long sentences using WordNet [33].

III. Semantic Representation of a Sentence

It is still impossible to fully represent the semantic elements of a sentence or text in AI. Geoffrey Leech suggests seven types of meaning, namely "conceptual, connotative, social, affective, reflective, collocative, and thematic", in his book "Semantics: The Study of Meaning" [34]. Semantic features depend on the meaning of words, word relationships, their position in the whole context (contextual features), their emphasis, references to the physical world such as color, time, geological location, natural laws, rhetoric, and even the understanding of the reader [35], [36].

In semantics, the concept is about "What is the text or sentence about and what does it refer to?". These are also the first questions we ask when we try to understand a text. Once we know the concepts, we can move on to the important relations, attributes, orders, and references. Before determining the similarity score in our study, we determined the nouns, verbs, and adjectives in the sentences and made a list for each one, as shown in Table I.

TABLE I. Preprocessing

Sentence 1: A woman is dancing and singing with other women. **Sentence 2:** A girl is dancing and singing in the rain.			
	Noun	**Verb**	**Adjective**
S1	Woman	Dancing, singing	Other
S2	Girl, rain	Dancing, singing	-

IV. Proposed Metric

Having presented each sentence as in the example in Table I, how can we determine whether or not the words in sentences are similar? It is not hard to see similarities if the words are not the same. We know from our daily lives that the human brain can understand the relationship between subordinate and superordinate words (hyponyms and hypernyms).

If we find the similarity value for each pair of words, we can average them to calculate the similarity between sentences. Equation (1) shows our proposed formula to calculate the similarity between sentences. We compare each word in one sentence with the words in the other sentence, and the most similar pairs of words are included in the calculation of the average.

$$sim_i = max_j \begin{cases} 1, & n_{1,i} = n_{2,j} \\ \dfrac{h_{cp,i,j}}{(h_{cp,i,j} - h_{1,i} + 1)(h_{cp,i,j} - h_{2,j} + 1)}, & n_{1,i} \neq n_{2,j} \end{cases} \quad (1)$$

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

$$sim_{i\_s} = max_j \begin{cases} 1, & n_{1,i} = n_{2,j} \\ \dfrac{h_{cp,i,j}}{(h_{cp,i,j} - h_{1,i} + 1)(h_{cp,i,j} - h_{2,j} + 1)(h_{max})}, & n_{1,i} \neq n_{2,j} \end{cases} \quad (3)$$

As can be seen in Figure 1, the starting node is the root element of the ontological tree; when we talk about the WordNet, the root word is "entity". $n_{1,i}$ is the i[th] word in the first sentence, and $n_{2,j}$ is the j[th] word in the second sentence. To calculate the similarity between $n_{1,i}$ and $n_{2,j}$, if $n_{1,i}$ and $n_{2,j}$ are the same, their similarity value is 1. For each i[th] element of sentence 1, we find the similarity value for all j[th] words in sentence 2, and the maximum similarity is considered.

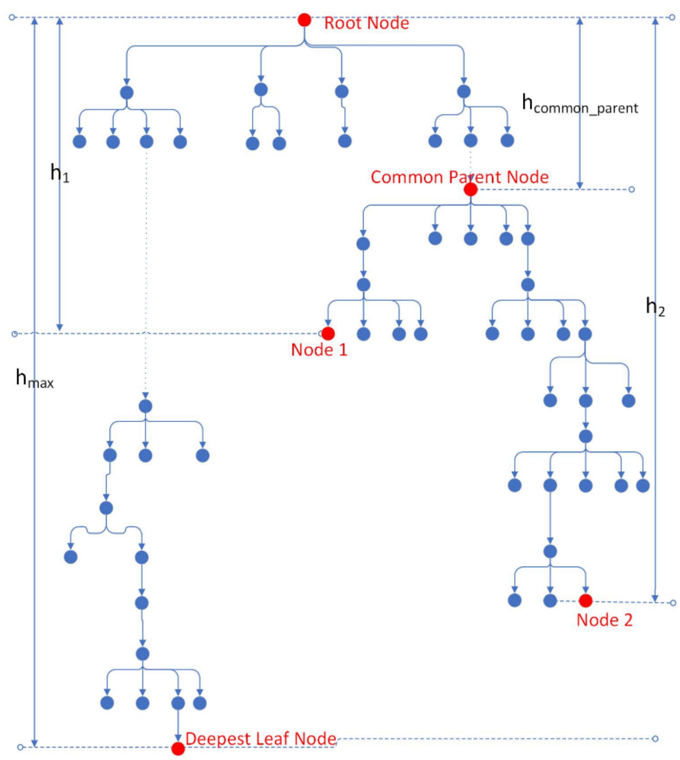

Figure 1. Nodes and their heights.

If the nodes are not equal, the distance is correlated with the height difference of the nodes to the common parent (n_{cp}). If both children are closer to the common parent, it means that the concepts of the children's nodes are also closer and similar. When the distance to the common parent is small, the similarity is high.

 ISBN: 978-1-68558-108-4

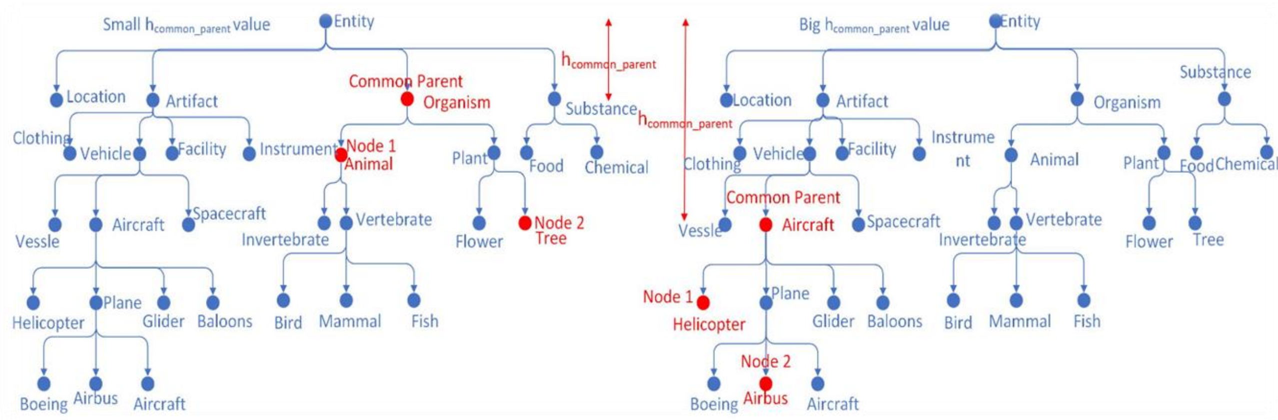

Figure 2. The depth of the common parent effect on similarity.

In the formulas given, $h_{cp,i,j}$ is the depth of the common parent node for the i^{th} word of sentence 1 and the j^{th} word of sentence 2. When we fix the common parent-child distance, the children are dissimilar when the common parent is closer to the root element. When a common parent is closer to the leaf node, its children are more similar, as shown in Figure 2. Consider the entity node. It initially has two children, one living and one non-living. However, when we go to the deepest nodes in the ontology, the two child nodes of “motorcycle” become more similar. They could be, for example, “motor scooters" and “mopeds".

Since the minimum similarity value of (1) is equal to zero and the maximum similarity value of (1) is equal to h_{max}, using min-max normalization yields the normalized similarity value as in (3). Figure 2 shows the depth of the common node effect in a simplified version of WordNet. In the left block, the depth of the common parent node is large compared to the root node. The parent node is closer to the leaf nodes than the nodes in the right block. Thus, in the left block, the children are more similar than in the second block, even though the depth difference between children and common parent is the same for the two examples in the right and left blocks. We can say that the depth of the common parent is inversely proportional to the similarity of the word.

The similarity of two different sentences is calculated using (4). n is the total number of features and similarity is the distance for each feature. If some words do not have a pair, the average similarity value decreases when divided by the number of words. We also consider the synonym-sets because a word may have more than one meaning and we take the average to decrease the error.

$$sim = \frac{1}{n}\sum_{i=1}^{n}(sim_i) \qquad (4)$$

Our algorithm is as follows:

- Step 1: For each pair of sentences in the dataset, remove the stop-words and the punctuations.
- Step 2: For each sentence in a pair, extract the Part-Of-Speech (POS) tags of each word.
- Step 3: Create a combination of the words in the pair according to their POS tags, then calculate the similarity score of the word pairs, using (3).
- Step 4: From the previous step, we have many similarity scores for a word. Accept the maximum similarity score.

V. DATASET & EVALUATION

We used the train split of the English STS benchmark dataset [37] to evaluate our proposal for computing semantic similarity. This dataset is a collection of data given in SemEval tasks between 2012 and 2017. It contains sentence pairs and their similarity scores. There are 5749 sentence pairs in the train split. The given similarity scores range from 0 to 5, where 0 means that the pairs have no similarity, and 5 means that the pairs are equally similar. These scores are annotated by human judges. To increase readability, we normalized the similarity scores using the min-max normalization function of scikit-learn [38].

We computed the pair similarity scores of the train split of the English STS benchmark dataset. First, the similarities are computed by considering only the nouns in the sentences. Second, the similarities are computed by considering both nouns and verbs in the sentences. For calculating the depth of the nodes in the taxonomy, we used WordNet. WordNet is a large electronic lexical database for English that proposes a hierarchical structure of concepts, where lower elements inherit information from their parents [22].

The similarities between the nouns in the sentences and the similarities between the verbs are averaged at the end to calculate the final similarity of the pairs. At last, the similarities are calculated by considering the nouns, verbs, and adjectives in the sentences. Again, the similarities between the nouns of the sentences, the similarities between the verbs, and the similarities between the adjectives are averaged at the end to calculate the final similarity of the pairs.

TABLE II. NORMALIZED PAIR SIMILLARITY SCORES EXAMPLES

Pairs	Sentences	STS Similarity Score	Noun Only Similarity Score	Noun + Verb Similarity Score	Noun + Verb + Adjective Similarity Score
1	A woman is dancing and singing with other women. A girl is dancing and singing in the rain.	0.60	0.73	0.86	0.56
2	Two men are packing suitcases into the trunk of a car. The men are putting suitcases into the car's trunk.	0.88	1.00	0.75	0.50
3	The woman picked up the kangaroo. A woman picks up a baby kangaroo.	0.75	1.00	0.50	0.33
4	Two foxes are eating from a plate on a brick patio. Foxes are eating from a plate.	0.56	0.51	0.75	0.50
5	Two zebras are playing. Zebras are socializing.	0.85	1.00	0.50	0.34
6	A group of people dance on a hill. A group of people are dancing.	0.64	0.67	0.33	0.22
7	A car is moving through a road. A car is driving down the road.	0.80	1.00	0.50	0.33
8	The man is shooting an automatic rifle. A man is shooting a gun.	0.76	0.58	0.79	0.52
9	A woman is cutting up a chicken. A woman is slicing meat.	0.55	0.54	0.27	0.18
10	Butter is being put into a bowl. A man cutting butter into a mixing bowl.	0.85	1.00	0.50	0.33

If any of the sentences of the pairs do not contain adjectives, then the similarity between the adjectives is zero. Therefore, the similarity score in such a case is drastically lower when we consider the similarity of the adjectives. In Table II, we have given ten sentence pairs with their normalized STS similarity scores as well as the similarity scores we calculated.

As can be seen from Table II, finding the nouns in the sentences and calculating the similarity of these nouns according to the proposed formula yields a meaningful similarity criterion.

Our similarity criterion is based on conceptual knowledge. As Lawrence W. B. Barsalou said, “The human conceptual system contains people's knowledge of the world. Conceptual knowledge in the conceptual system supports a variety of basic cognitive operations, including categorization, inference, and the representation of propositions.” [1].

To get an idea of how the proposed method works with the dataset, we chose different threshold values. When the similarity threshold for a set of normalized similarity values is set to 0.50, any value greater than or equal to 0.50 is considered similar.

For these sets, we examined what percentage of these sentence pairs had a similarity of 0.50 or greater. Similarly, we chose the threshold value of 0.25 to understand the percentage of pairs that are not strongly similar and not strongly dissimilar. In Table IV, we have given the percentage results for different word selections and threshold settings.

During the evaluation, we performed the following tests for nouns, noun+verb, and noun+verb+adjective. We used the train split of the English STS benchmark dataset, which has normalized similarities between 1.00–0.75, 0.75–0.50, and 1.00–0.50, and checked what percentage of pairs the proposed method finds in this range to see if our proposed method also labels these pairs similarly. Again, we used the same data set, which has normalized similarities between 0.50 – 0.25, 0.25 – 0.00, and 0.50 – 0.00, meaning that the pairs are not similar. We checked what percentage of pairs the proposed method finds in this range to see if our proposed method also names them similarly.

Our Pearson correlation results can be found in Table III. Here, we measured the correlation between the normalized similarity values of the data set and the normalized similarity values we calculated.

TABLE III. CORRELATION RESULTS

Task Name	Pearson Correlation Score
Noun Only	0.51
Noun + Verb	0.47
Noun + Verb + Adjective	0.48

TABLE IV. EVALUATION

	Results of the Proposed Method	
Noun Only	Percentage of pairs in the range of 1.00– 0.50	77.09%
	Percentage of pairs in the range of 1.00 – 0.75	72.42%
	Percentage of pairs in the range of 0.75 – 0.50	82.05%
	Percentage of pairs in the range of 0.50 – 0.00	56.85%
	Percentage of pairs in the range of 0.50 – 0.25	42.38%
	Percentage of pairs in the range of 0.25 – 0.00	69.45%
Noun + Verb	Percentage of pairs in the range of 1.00 – 0.50	56.58%
	Percentage of pairs in the range of 1.00 – 0.75	50.74%
	Percentage of Pairs in the range of 0.75 – 0.50	62.77%
	Percentage of pairs in the range of 0.50 – 0.00	71.94%
	Percentage of pairs in the range of 0.50 – 0.25	59.37%
	Percentage of pairs in the range of 0.25 – 0.00	82.88%
Noun + Verb + Adjective	Percentage of pairs in the range of 1.00 – 0.50	46.35%
	Percentage of pairs in the range of 1.00 – 0.75	19.47%
	Percentage of pairs in the range of 0.75 – 0.50	74.88%
	Percentage of pairs in the range of 0.50 – 0.00	81.99%
	Percentage of pairs in the range of 0.50 – 0.25	54.76%
	Percentage of pairs in the range of 0.25 – 0.00	95.71%

We ran our experiments on an Intel(R) Core(TM) i7-9750H CPU. The entire experiment took about 48 seconds.

VI. CONCLUSION

In this study, we have proposed a formula for calculating the conceptual similarity of sentences. In the formula, the coefficients calculate how similar the noun words that make up the sentence ("verbs" and "adjectives" are also included) are to their most similar conjugates in the other sentence by looking at the distance of these two words to their common ancestor and the location of the common ancestor in the ontology tree. If the compared words are close to their common ancestor, they are more likely to be similar. The other important parameter is the depth of the common ancestor in the ontology tree. If the common ancestor is far from the root, the similarity of the compared words increases according to its position closer to the root node.

Since we are interested in the conceptual similarities, even if WordNet also has a taxonomic structure of adjectives that indicate the attribute of the nouns (concepts), they are not the actual concepts [39]. The inclusion of the similarity contribution between verbs or adjectives in our study negatively affected the results. This may be understandable if we consider similarity as a conceptual method.

In our upcoming research, we want to use a dataset where the conceptual similarity of sentences is scored by humans. To decide on a 5-level similarity scale (high similarity, low similarity, different, completely different, and no idea), participants are asked to use crowd-sourcing methods. How meaningful the results are determined by comparing the proposed similarity calculation with the human markers. We expect our method to give better results on the human tagged datasets, since our proposed method simulates the human mind to find the conceptual relationship.

REFERENCES

[1] L. W. Barsalou, “The human conceptual system,” The Cambridge handbook of psycholinguistics, pp. 239-258, 2012.

[2] G. Pirró and D. Talia, “UFOme: An ontology mapping system with strategy prediction capabilities,” Data & Knowledge Engineering, vol. 69(5), pp. 444-471, 2010.

[3] F. Benedetti, D. Beneventano, S. Bergamaschi, and G. Simonini, “Computing inter-document similarity with context semantic analysis,” Information Systems, vol. 80, pp. 136-147, 2019.

[4] J. Nasir, I. Varlamis, A. Karim, and G. Tsatsaronis, “Semantic smoothing for text clustering,” Knowledge-Based Systems, vol. 54, pp. 216-229, 2013.

[5] W. Song, J. Liang, and S. Park, “Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering,” Information Sciences, vol. 273, pp. 156-170, 2014.

[6] S. Kumar and K. Bhatia, “Semantic similarity and text summarization based novelty detection,” SN Applied Sciences, vol. 2, pp. 1-15, 2020.

[7] A. Vretinaris, C. Lei, V. Efthymiou, X. Qin, and F. Özcan, “Medical entity disambiguation using graph neural networks,” In Proceedings of the 2021 International Conference on Management of Data, pp. 2310-2318, 2021.

[8] V. Demertzi and K. Demertzis, “A hybrid adaptive educational eLearning project based on ontologies matching and recommendation system,” arXiv preprint, arXiv:2007.14771, 2020.

[9] A. Chikkamannur, “Semantic Annotation of IoT Resource with ontology orchestration,” In 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC- IEEE) pp. 1-7, 2020.

[10] S. Mhammedi, H. El Massari, and N. Gherabi, “Composition of large modular ontologies based on structure,” In Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C’21, pp. 144-154, 2022.

[11] A. Belhadi, Y. Djenouri, G. Srivastava, and J. Lin, “Fast and Accurate Framework for Ontology Matching in Web of Things,” ACM Transactions on Asian and Low-Resource Language Information Processing, 2023.

[12] G. Yang et al., “CCGIR: Information retrieval-based code comment generation method for smart contracts,” Knowledge-Based Systems, 237, 107858, 2022.

[13] M. Sreenivasan, S. Dhar, and A. Chacko, “PCPS: Personalized Care through Patient Similarity,” In 2022 IEEE Region 10 Symposium (TENSYMP-IEEE), pp. 1-6, 2022.

[14] Y. Liu and M. Ijaz, “Personalized auxiliary information presentation system for mobile network based on multimodal information,” Mobile Networks and Applications, pp. 1-11 2022.

[15] F. Liu and S. Li, “Research on personalized user-centered product improvement based on sentiment mining of online reviews and competitor analysis,” Home, http://www.researchsquare.com/article/rs-1829215/v1 (Accessed Aug. 19, 2023).

[16] S. Babalou, A. Algergawy, and B. KönigRies, “SimBio: Adopting Particle Swarm Optimization for ontology-based biomedical term similarity assessment,” Data & Knowledge Engineering, 102137, 2023.

[17] M. Landolsi, L. Hlaoua, and L. Ben Romdhane, “Information extraction from electronic medical documents: state of the art and future research directions,” Knowledge and Information Systems, vol. 65(2), pp 463-516, 2023.

[18] B. Yang et al., “Classification of Medical Image Notes for Image Labeling by Using MinBERT,” Tsinghua Science and Technology, vol. 28(4), pp. 613-627, 2023.

[19] D. Tian, M. Li, Y. Shen, and S. Han, “Intelligent mining of safety hazard information from construction documents using semantic similarity and information entropy,” Engineering Applications of Artificial Intelligence, vol. 119, 105742, 2023.

[20] N. B. Cam, Sementic Similarity. Github. https://github.com/bengisucam/semanticSimilarity . (accessed Aug. 19, 2023).

[21] R. P. Honeck, “Semantic similarity between sentences,” Journal of Psycholinguistic Research, vol. 2, pp. 137-151, 1973.

[22] G. A. Millar, “WordNet: A Lexical Database for English.,” In Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

[23] M. Sussna, “Word sense disambiguation for free-text indexing using a massive semantic network,” In Proceedings of the second international conference on Information and knowledge management, pp. 67-74, 1993.

[24] J. J. Jiang and D. W. Conrath “Semantic similarity based on corpus statistics and lexical taxonomy,” arXiv preprint cmp-lg/9709008, 1997.

[25] N. Seco, T. Veale, and J. Hayes, “An intrinsic information content metric for semantic similarity in WordNet,” In Ecai vol. 16, p. 1089, 2004.

[26] D. Yang and D. M. Powers, “Measuring semantic similarity in the taxonomy of WordNet,” Australian Computer Society, 2005.

[27] X. Y. Liu, Y. M. Zhou and R. S. Zheng, “Measuring semantic similarity in WordNet,” In 2007 international conference on machine learning and cybernetics, vol. 6, pp. 3431-3435, 2007.

[28] Z. Zhou, Y. Wnag, and J. Gu, “New model of semantic similarity measuring in wordnet,” In 2008 3rd International Conference on Intelligent System and Knowledge Engineering, vol. 1, pp. 256-261, 2008.

[29] P. Sravanthi and B. Srinivasu, “Semantic similarity between sentences,” International Research Journal of Engineering and Technology (IRJET), vol. 4, no. 1, pp. 156-161, 2017.

[30] A. Selvarasa, N. Thirunavukkarasu, N. Rajendran, C. Yogalingam, S. Ranathunga, and G. Dias, “Short Tamil sentence similarity calculation using knowledge-based and corpus-based similarity measures,” In 2017 Moratuwa Engineering Research Conference, pp. 443-448, 2017.

[31] SemEval 2016 Dataset: https://altqcri/semeval2016/task2. (Accessed Aug. 19, 2023)

[32] M. N. Jeyaraj and D. Kasthurirathna, “Mnet-SIM: A multi-layered semantic similarity network to evaluate sentence similarity,” International Journal of Engineering Trends and Technology, vol. 69, no. 7, pp. 181–189, 2021. doi:10.14445/22315381/ijett-v69i7p225

[33] M. C. Lee. “A novel sentence similarity measure for semantic-based expert systems,” Expert Systems with Applications, vol. 38, no. 5, pp. 6392-6399, 2011.

[34] N. Love, “Translational semantics: A discussion of the second edition of Geoffrey Leech's Semantics: The Study of Meaning,” Stellenbosch Papers in Linguistics, vol. 11, pp. 115-136, 1983.

[35] D. Geeraerts, “Theories of lexical semantics,” OUP Oxford, 2009.

[36] T. A. Van Dijk, “Society and discourse: How social contexts influence text and talk,” Cambridge University Press, 2009.

[37] Cer, Daniel et al., "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," arXiv preprint arXiv:1708.00055, 2017.

[38] Pedregosa, Fabian et al., "Scikit-learn: Machine learning in Python," The Journal of machine learning research 12, pp. 2825-2830, 2011.

[39] D. Gross and K. J. Miller, “Adjectives in WordNet,” International journal of lexicography, vol. 3, no. 4, pp. 265-277, 1990.